

Data Fraud Detection: A First General Perspective

Hans-J. Lenz^(✉)

Institute of Statistics and Econometrics, Freie Universität Berlin,
Boltzmannstr. 20 K30, 14195 Berlin, Germany
hans-j.lenz@fu-berlin.de

Abstract. We try to present a first broad overview on data fraud, and give hints to data fraud detection (DFD). Especially, we show examples of data fraud that happened at anytime of human mankind, all around the world, and affects all kind of human activities. For instance, betrayers are entities of the society, industry, banks, services, health-care, non-profit organizations, art, science, media or even a government or the Vatican. We consider four main areas of data fraud: spy out, plagiarism, manipulation and fabrication of data.

Of course, there is not only interest on data fraud itself but on its detection, too. Although improvements of data fraud detection is evident, it seems that the intellectual creativity and capacity of the betrayers is unlimited. Especially, the Internet with its various services and the mobile communication opened the Pandora box for criminal acts. Furthermore, one may state the hypothesis that while the ethics behavior of people decreases over time the data fraud rate is continuously increasing.

There does not exist an omnibus data fraud detector, and the author supposes there will be never one upcoming due to the heterogeneity of the domain. For instance, compare the domains “spy out” in industry and “data fabrication” of observational or experimental studies in science. It is a matter of fact that the interest and need of science, business and governmental authorities is increasing over time for improving tests of data fraud detection. This paper can be viewed as a modest attempt for stimulating research into this direction.

1 Introduction

One can best unfold the complexity of data fraud by classifying fraud into four classes. All of them are driven by three time-invariant features of the societies, business and human beings: Power, glory and money. As saying goes “Knowledge is Power” power is the driving force of **Spy out** in the military and secret service area, while knowledge and profit(money) are the main factors in business. The activities of the secret services of all countries like *Central Intelligence Service (CIA)* or *NSA* in USA, *Military Intelligence (MI6)* in UK, the *Russian secret service (CBP)* etc. go back far into history. No doubt, they have ever had influenced failure and success of military actions at any time. Less power and glory but more profit (money) has been the forcing power of industrial or business

spy out. There is no sharp boundary to the extensive “silent” storing and analytic analyses of customer’s data, mostly without explicit permission. Consider only the massive accumulation of such data by *Amazon*, *Ebay* or *Google*. Together with nearly unlimited **memory** and **Big Data Analytics** it demonstrates the increasing risk of leaving *Recommendation Systems*, simple pull/push systems, and entering into a *Total Information World* with loss of any privacy of anybody, anywhere and anytime as an extension of *ubiquitous computing*.

The second domain of data fraud is related to **Plagiarism**. It concerns mainly the illegal usage of data of somebody else, sometimes existing as *self plagiarism*. Since the start of the digital era plagiarism M.Sc. or Ph.D. studies have gained more and more attention, although plagiarism is not limited to master or doctoral studies, cf. plagiarism happening in composing, drawing and painting. As we shall see later the main interest in *plag* detection in the domain *PhD dissertations* is caused by spectacular cases and the development of improved methods (citation based *plag* detection besides of substring matching), [4], and “swarm intelligence”. The last approach makes use of the efficiency of many, more or less independently acting *plag hunters* pooling information about the same entity.

Our main interest, however, is devoted to **Data Manipulation**. Its main characteristic is given by the dishonest manipulation of existing own or foreign data. The objective is gaining prestige or making money. Evidently, data manipulation happened at any time, anywhere and every domain of life is influenced. For instance, the annual fraud rate of the new U.S. health care system is estimated to be about 10%, [29]. Consider the scandalous manipulation of clinical trials at the Medical University of Innsbruck (MUI), Austria, by Dr. H. Strasser, [19], or the Libor 3(5)-months interest rate manipulation jointly performed by a cartel of *Deutsche Bank*, *Royal Bank of Scotland (RBS)*, *Union Bank of Switzerland (UBS)*, and *Barclays Bank*, UK, [20]. Greece fudged its annual depths-GDP rates in the years before the country applied for entering the Euro zone, [22]. In recent times the dishonest shuffling of ranks for TV shows by non-profit organizations like ADAC, [23], or ZDF, Germany, [24], or even black money transfers and money laundry by the Vatican, [25], caused much attention in the media.

A question of historical interest remains to be answered later: Who is known to be the first betrayer in science? It is a tragedy of science that that profession is more and more inclined to manipulate data gained from empirical (observational or experimental) studies. The causes are simple the pressure caused by the dominating principle “Perish or Publish”, the increasing know-how needed for applying sophisticated methods of data science, and the need of fund raising for ongoing research. Yet, there exists a further variant of data fraud which is far beyond the “simple manipulation” we talked about.

Data Fabrication. seems to be the most awful or “mostly criminal” form of data fraud. Instead of manipulating data the betrayer selects an easier way of collecting data: He simply generates the artificial data he needs. Evidently, this saves cost and time, and, consequently, increases the chances of fund raising or early publication because application forms or research papers can be submitted with some time lead. Furthermore, the data will perfectly fit the betrayer’s

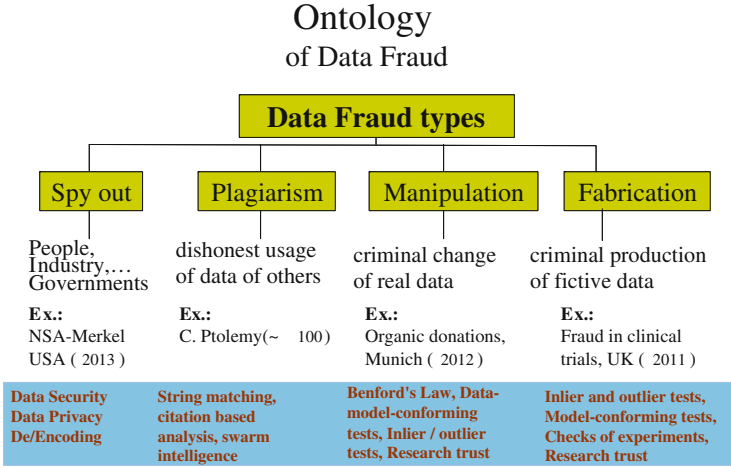


Fig. 1. Data fraud types, examples and detectors.

hypothesis or objective if “professionally” created. A shocking case of such academic data fabrication is related to the Dutch psychologist D. Stapel, Univ. of Tilburg, [21] whose forgery finally was stopped in 2013.

In the following we shall give a short history of spectacular cases of data fraud. We pass by spying out, and turn to plagiarism with a special emphasis on doctoral dissertations. We deeply look at data manipulation, fabrication and the corresponding detection techniques. From the methodological point of view the domains *manipulation* and *fabrication* strongly overlap. Consequently, the same methods of detection can be used, i.e. **Statistics**, **Data Mining**, and **Machine Learning**. Today, that bundle is labeled **Data Analytics**.

An overview on the four types of data fraud is given in Fig. 1 where we limit ourselves to one example per each fraud type, and list some fraud detectors at the bottom line.

2 A Short History of Data Fraud

To best of the author’s knowledge the first spectacular case of data fraud (data plagiarism) happened in ancient Egypt. **Claudius Ptolemy** (85 – 165 p. C.) was a leading astronomer and mathematician at Alexandria, the center of the antic world of science. He was the father of the *Almagest*, the famous Arabic star calendar. He used the astronomical data from **Hipparchos of Nicaea** (190 – 120 a. C.), but he did not refer to him, [6]. Historians may argue here that the current point of view of referencing might not be adequate for that science period. Today such kind of fraud is called **Data Plagiarism**. The lesson for detecting such a fraud is to carefully **Check the Provenance** of observational data referred to in a publication.

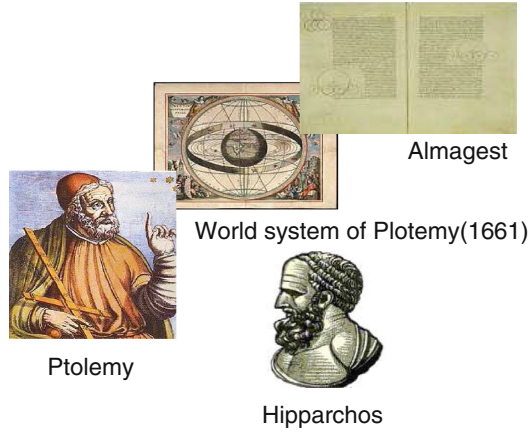


Fig. 2. Ptolemy, his centric world view, the Almagest, and Hipparchos, images: [10–13].

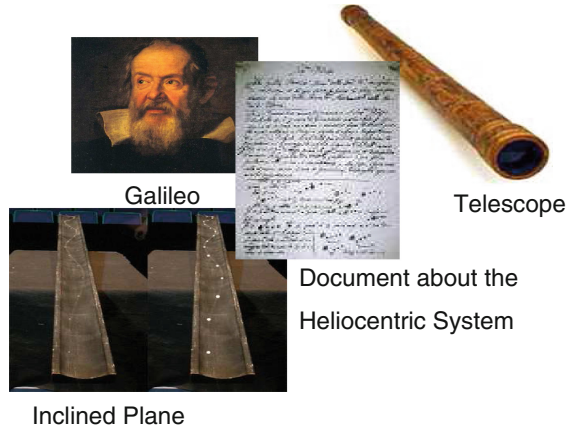


Fig. 3. Galileo, his heliocentric system, telescope and plane model, images: [14–17].

There is no doubt that *Galileo Galilei* (1564 – 1641) was one of the giants in science. For example, he made ingenious contributions to astronomy and physics. Especially, his astronomical research on the heliocentric system, and his experiments with bodies sliding down inclined planes are famous. At that time, the devices for measuring run time of moving objects produced “large” measurement errors. Galileo believed in the validness of his cinematic hypotheses, and for supporting them he decided to reduce the measurement errors. Today, there is no doubt that Galileo’s experiments could not have been run in its way. As [7, 176] stated “The Genius was motivated by the objective of supporting the final break-through of his ideas.” Refer to [8, 27] for further details. Simply speaking, motivated by prestige Galileo used data manipulation thus violating the **Principle of Reproduceability** of well-performed experiments.



$$F_{\text{net}} = m \cdot a$$

$$F_{\text{grav}} = m_1 \cdot m_2 / d^2$$

$$\dots$$

Fig. 4. Newton “Cuius genius humanibus superavit”, image: [18].

A further hero of science was *Sir Isaac Newton* (1643 - 1726) whose scientific contributions to astronomy, mechanics (acceleration, gravitation and forces) and mathematics (Calculus) opened a new era of physics, especially kinematics. In his book *Principia* Newton convinced due to the reported high precision of his observations far from being legitimate, [7,9]. Or as [7, 176] or [9, 1118] put it “Nobody was so brilliant and effective in cheating than the master mathematician.” Newton suppressed the real imprecision of his measurements being anxious about a non perfect fit and casting doubts on his theory in the scientific community. In order to support his hypotheses he faked the output of his experiments and the observations by downsizing the errors. Accordingly, this manipulation contradicts the *Principle of Reproduceability*. His motive certainly was not profit, but fear of loss of prestige.

Data manipulation has many variants as we shall see later. One form is the trick used by experimenters to select and publish only a proper subset of results or runs. A “representative” case is given by the physicist *Robert Millikan*, US Nobel Price Winner, 1868 – 1953. The claim is his strikingly precise measurements of the charge of electrons in 1913 being quite better than those of his rival *Felix Ehrenfeld* who experienced large deviations, [7,8]. Milikan’s lab protocols showed later that he published only the ‘best 58 out of 140’ experiments having smallest measurement variance, [8,34] and [7, 176–177]. Today such a misbehavior of running experiments is called **Experimental Selection Bias** and contradicts, of course, the *Principle of Reproduceability*. Here again *Prestige* combined with Nobel price winning expectations was presumable the driving force of data fraud.

We continue presenting historical cases of data fraud and turn to the *Deutsche Bank* as an example from the business sector. Alternatively, we could have selected from the U.S. economy the *Enron* case with falsification of balance sheets, [47]. *Josef Ackermann*, born 1948, was the chairman of the board of directors 2002–2012, and was responsible for all claims. Under his leadership the *Deutsche Bank* was involved in many scandals like manipulating the prices of

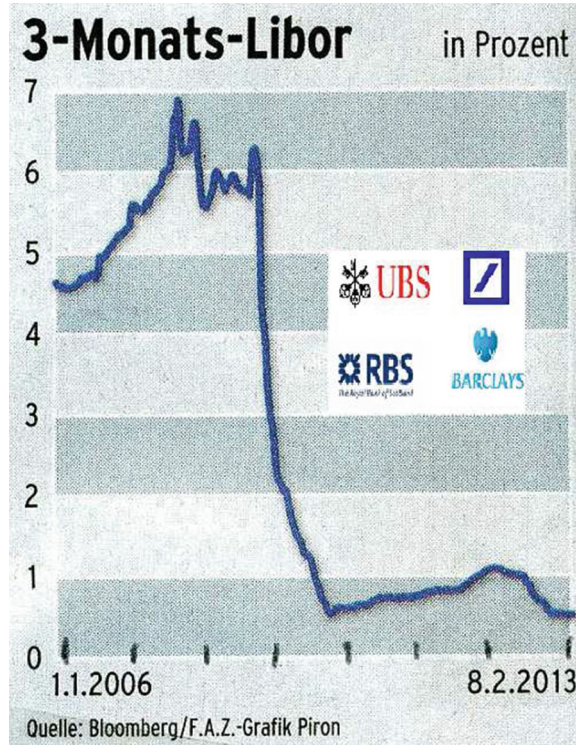


Fig. 5. Libor manipulation by the cartel, image: [20].

food products, unlawful CO_2 -emission permits and financial derivatives trading, and, last but not least, manipulating the 3(6) months Libor interest rate within a bank cartel consisting of Barclays, Deutsche Bank, RBS and UBS, [20]. The legal task of the group was a “fair price fixing” of both interest rates, but the cartel used its economic power for down or upraising of the Libor for making moderate but continuous profits, cf. Fig. 5.

Another domain where forgery typically happens is health care, cf. [29] for fraud detection in the new U.S. medicare system. Especially medical universities are affected institutions because of the pressure to economically manage clinics, fund raising for research or corruption. The last motive was underlying extensive sport doping and manipulation of facts and figures at the Univ. of Freiburg, Germany, cf. [27]. Above we already mentioned the case of illegal experiments at the *Medical University of Innsbruck (MUI)*, Austria, as reported by *Nature*, [19]. In Germany tricksters manipulated the lists of organic transplantations at several of the 49 German transplantation centers like Göttingen, München, Münster, Regensburg etc. [26]. In order to detect fraud in such cases independent experts are employed who checked with diligence all mails, lab protocols, lists, reports, time-schedules and revenues and expenditures vouchers with respect

to facts and prior knowledge in a qualitative and quantitative way. We call these (qualitative) manual activities assisted by analytical techniques **Manual Inquest of Data Fraud**. They are based on **Abductive Reasoning**, and are similar to trouble shooting from an investigation point view.

Next, we turn to economics. In this domain data fraud happens for gaining more economic influence or reward. As an example take Greece in the late nineties. The crucial period for Greece to enter the Euro zone was 1997 – 1999. Greece manipulated its official macro-economic statistics, and made the EC member countries, the EC authorities in Bruxelles and Luxembourg and the public believe that the deficit had fallen under the *Maastricht limit* of 3%. However, later the (new) Greece Finance Minister, Mr. *Aligorskoufis*, confessed: “It has been proven that the deficit had not fallen below 3% in every year since 1999.” [22]. Indeed, the figures for 1997-1999 were 6.44, 4.13 and 3.38%.

Let us close with politics and consider governments who do not completely obey democratic rules. It is a matter of fact that the elections in Russia 2011 were fudged, [28]. Especially, the result of *Putin’s* party *United Russia* is doubtful. Consider the empirical distribution of electoral votes on the percentages of votes in the interval $[0, 100]$. Typically, such a distribution is bell-shaped and is smoothly curved in the center and at both tails. Due to the evident manipulation of votes the distribution of *Putin’s* party is heavily skewed to the right (large percentages), and shows small up-and-down spikes above 30% around the values 35, 40, 45, . . . , 100%, cf. Fig. 6.

The lesson learned from the last type of data manipulation that sometimes simply plotting of histograms (empirical distributions) is an effective “starter” for further investigations on suspectable data.

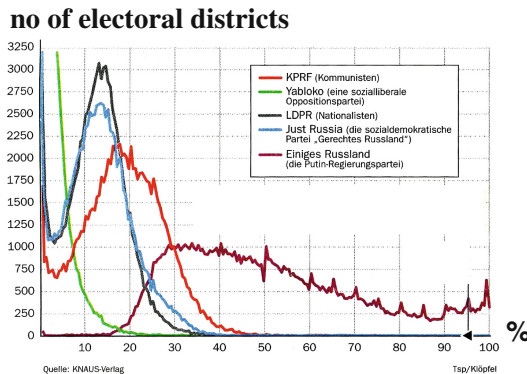


Fig. 6. Manipulation of Russian elections in 2011, image: [28] layout modified.

3 Spy Out

One can raise the question whether or not **Cheating** by copying text or formulas from neighbors during examinations at school is an entry into the world of spy

out or plagiarism, and is abnormal or not. No doubt that kind of spying out is happening at a very low fraud level.

From a historical perspective massive **Spy out** is related to ancient and modern populations and their military actions against other countries. It may be considered as a prerequisite of any war and struggle, and aims for getting better information about the enemy for improving the success of their own attacks. Think only of the empires Greek and Persia, Egypt and Hittites, or Roman and Germanic. The secret services of all modern countries around the globe are expected to deliver information about the rest of the world. After 9/11 the National Security Agency (NSA), USA, has scaled-up its computing facilities for spying out everybody's data on an extreme large scale. The question arises whether "Big brother is watching you" is realized or not. Here any kind of world-wide communication by phone, email, Twitter or any other media is affected. Even encrypting is not at all safe and a guarantee for people's privacy. The fact that even the German chancellor's telephone is tapped by US authorities is a nasty perspective.

Let us leave this domain and turn to economics and business. Telephone fraud became roughly a problem at the end of the eighties. Legal telephone cards with or without a credit were stolen and manipulated by betrayers which made phone calls possible free of any charge. Corruption related to data fraud was recently reported by the German Press Agency and printed in *Der Tagesspiegel*, a leading newspaper in Berlin, Germany. Employees of the sanctuary health assurance company *Debeka* were accused of having tried to get data of state employees candidates from corruptive state employees targeting for new contracts, [5].

As buying and selling is an intrinsic part of daily life of consumers, tricksters and betrayers consider trade as their domain for data fraud. Each trade has a phase "selecting and putting an item into the shopper basket" and "paying". Paying is mostly done by cash, alternatively by EC or credit cards. Forget cyber money here due to its minimal market share. **Card Fraud** is concerned with card theft or the misuse of bank account and PIN data. Therefore, we have the following four classes of fraud, cf. [2]:

- Theft
- Duplicate Generation
- Skimming
- Phishing.

Theft is a real fraud if a credit card like VISA, AMERICAN EXPRESS or MASTER is stolen and used for money withdrawals. However, when a card theft is combined with **Card Cloning** (in the sense of duplicate production) things become more tricky and dangerous, and increase the card owner's risk of big monetary losses. The thief simply can withdraw money from various sites. **Skimming** starts at the other end of the buying chain just to say so. Before a consumer pays by cash he has to collect money, for instance, from a teller machine (ATM). Tricksters install a small web cam and a thin keyboard at the ATM's site for catching the bank account or credit card number together with

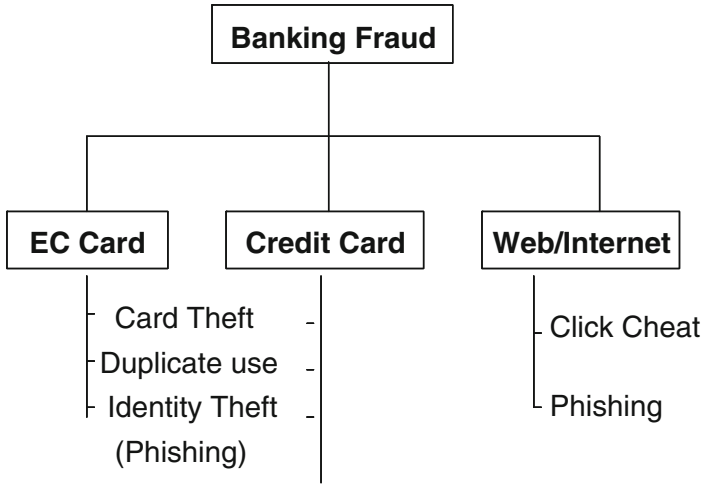


Fig. 7. Fraud of banking and internet banking.

the PIN from an innocent client. **Phishing** refers to sending faked emails to users with a subject like “Full mail box” or “Special offer”, and asking them for delivering their account, password (PIN) and transaction number (TAN). In a similar way, novice surfers are redirected to false (criminal) links or attracted by special (low priced) offers for visiting specially prepared web pages. When such sites are entered the trickster fishes the bank account, PIN or name, credit card number and safety code for starting his own criminal transactions. Let us add **Advertising Click Fraud** where *Pay per click* systems are manipulated by artificially generating clicks by the betrayer for illegally making money at the expense of the advertising company, [2].

For an overview by a diagram of data fraud related to card usage or Internet banking have a look at Fig. 7. There exists a big bundle of **Multivariate Explorative Data Analysis Techniques** for detecting EC or credit card fraud, especially methods like *Generalized Linear Regression* and *Classification*.

4 Plagiarism

As noted above **Plagiarism** is considered to be the dishonest use of data of a second party. It has many facets and exists in many fields of human life like arts, business and science. The motives are either prestige or mostly profit. A short overview is given in the diagram in Fig. 8 below.

We present famous cases of plags: Pirating of documents (“Galileo Forgery”) and doctoral dissertation plag of the former German minister of defense, *K.-T. von Guttenberg*. Both cases happened in Germany, and the forgery was proven in 2012 and 2011, respectively.

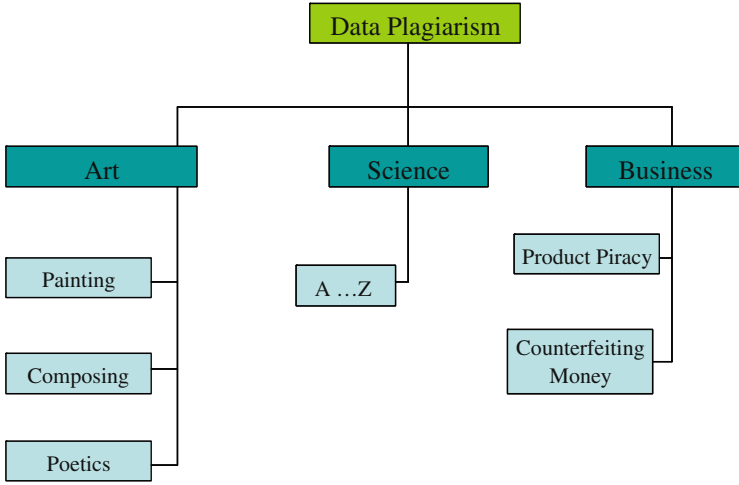


Fig. 8. Selected domains of plagiarism.



Fig. 9. Discrepancies between the original (“Sidereus Nuncius”) and faked book, image: A Galileo Forgery.

Galileo’s book “Sidereus Nuncius” was published in 1610, and it is a fact that 80 legal copies are known to exist worldwide. In 2005, surprisingly, a further (faked) copy including until then eighty unknown (faked) ink drawings was offered by art dealers on the international market. The German expert Horst Bredekamp, Humboldt Univ. Berlin, used the drawings from the faked copy in a book he published entitled “Galilei, der Künstler”. Finally, in 2012 a British historian proved the forgery done by the Italian M.M. De Caro. His detection caused Bredekamp et al. to edit the book “A Galileo Forgery” where he compiles the events and explains what happened. It is interesting to note that besides of various material analytical techniques **Pattern Matching** of symbols (stars) and characters (L) helped discriminating between the original and faked document, cf. Fig. 9.

One of the most spectacular data plagiarism claims corresponding to doctoral dissertations was caused by the work of K.-T. von Guttenberg, the former

minister of defense of Germany. Today his dissertation can be considered one of the best investigated plagiarism. It may be even considered as a yardstick for further improvements of academic plagiarism detection software. While most of the plag software uses simple **Substring Matching** techniques it was the idea of *B. Gipp* (2011), [4], to use a kind of *meta analysis* based on *information retrieval methods* for developing his **Citation based Plag Detection (CbPD)** technique. Several steps are built-in: Identify citation patterns by order, non-lexical proximity and distinctiveness of in-text citations. This idea enabled him to detect multi-lingual and multi-source detection plagiarism by cross-reference of the citation structures of an original and its plag. All computational output is carefully visualized. Especially, the hit rate of detecting disguised plag like paraphrases and cross-language plag could be clearly increased. It is a great advantage of Gipp's approach that his procedure produces higher confidence levels than the existing matching algorithms for academic plagiarism detection, [4,30]. Simple string matching or substring search work as follows, cf. the plag platforms *citeplag*, *Vroniplag* etc. Let \mathbf{A} be an alphabet, and $s_1 \in \mathbf{A}^m$ and $s_2 \in \mathbf{A}^n$ literal text strings where $m \leq n$. Prove $s_1 \subseteq s_2$ or $\text{sim}(s_1, s_2) > s_{low}$! We present a barcode representation of the thesis produced by manual inspection done by the members of the GuttenPlag Wiki project, see Fig. 10. The semantics of coloring is as follows:

- red bars: multiple sources plag pages
- black: single source plag pages
- white: no plag pages
- blue: content and bibliography pages.

It was found that 64 % of all lines of the text was plagiarized. More details of *v. Guttenberg's* plagiarism case can be unscrambled by using the plag location and visualization prototype *CitePlag* developed by Gipp(2013). The software offers five citation-based approaches using two documents as input - a plag candidate and the original - [30]. We present only one illustration here. In Fig. 11 the left image shows the dense citation patterns produced by the *Bibliographic Coupling* method in *v. Guttenberg's* thesis. The image to the right shows the concept of citation chunking. Numbers represent matching citations occurring in both documents, and the letter x indicates non-matches.

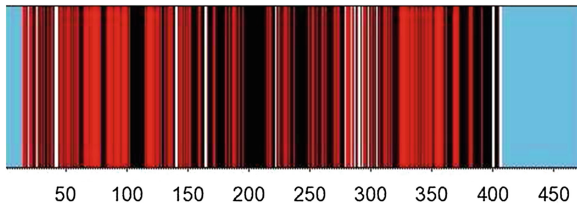


Fig. 10. Barcode of Guttenberg's thesis by plag type, image: [31].

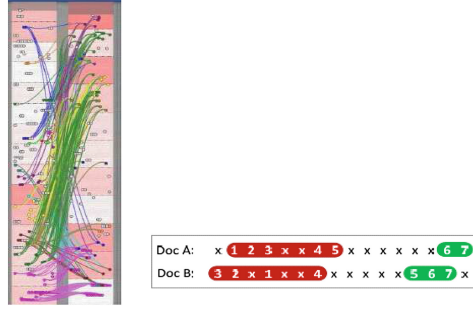


Fig. 11. Citation chunking of Gutenberg's thesis and its concept, images: [30].

- *Bibliographic Coupling* is a similarity measure between two reference lists
- *Citation Chunking* is a citation pattern matching irrespective of the order of matching citations
- *Greedy Citation Tiling* identifies longest citation patterns consisting entirely of matching citations in the exact same order
- *Longest Common Citation Sequence* searches for the longest string of citations matching in both documents in identical order
- *Longest Common Sequence of Distinct Citations* includes only the first occurrence of a matching citation. It ignores repeated citations of the same source regardless of occurring in the same order in both documents or not.

5 Data Manipulation and Fabrication

As mentioned above **Data Manipulation** and **Fabrication** overlap from a methodological point of view. Of course, the data generation process is quite different. Therefore it makes sense to present fraud detectors for both fraud kinds together.

Data manipulation is the dishonest change of the content of existing own or third party data, irrespectively, whether the content is encapsulated in text documents or not, i.e. tables, diagrams or photos. In most cases numbers are manipulated. Data fabrication is the criminal production of artificial figures driven by gaining power, prestige or profit (“Gier”).

In the following we mainly focus on data fraud in science. There exists a lot of studies on the various types of fraud. A recent field study based on interviews and questionnaires was published in *Nature* in 2005, and was authored by *Martinson, Anderson* and *de Vries*, [32]. They addressed $N = 7760$ scientists who got a grant from the National Institutes of Health, USA. The final sample size was $n = 3247$, i.e. the response rate is only slightly less the 42%. The field study was designed as an anonymous self-report based on a standardized questionnaire with $(10 + 6)$ items of strong and medium fraud types. The study differed between early and mid-career researchers. Figure 12 summarizes the quite disappointing aggregated results.

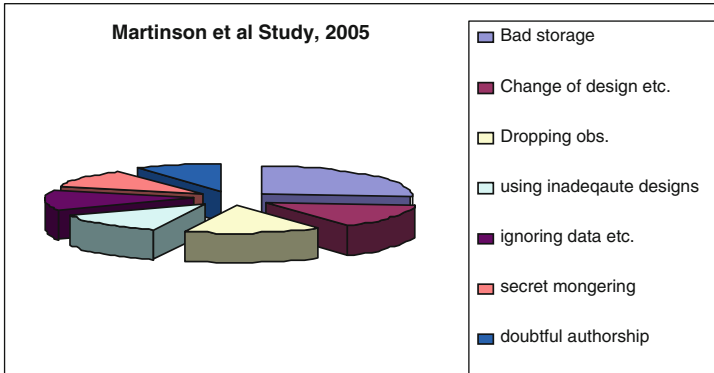


Fig. 12. Types of data manipulation, compiled in the Martinson et al. study, image: [32].

Empirical data is no longer accessible, designs are doubtful, and inappropriately application of statistical methods like sub-setting, sequential testing or manipulation of p-values. From the author's point of view the *Martinson et al.* report supports the conjecture of time-stability of such findings in psychology, social science and medicine. Further contributions supporting this hypothesis are due to J.P.A. Ioannidis and L. John, [33,34]. Again, remind yourself that data manipulation exists everywhere at any time in science. For instance, the Japanese stem-cell researcher H. Obokata, Kobe, was forced in 2014 to retract her co-authored papers in *Nature* because of four main allegations raised about her research, [35]:

- Irregularities of published images
- Identical text copied from another own paper without reference (self plag)
- Inconsistencies between published papers and later author's explanation
- Non Reproduceability of experimental results.

Next we present some few cases where statistical methods are not correctly used leading to surprising results. One has to confess that the borderline between non professional usage of statistics and data manipulation is fuzzy.

- Although the diastolic and systolic blood pressure are stochastical dependent, doctors use them independently of each other for their diagnoses worldwide.
- Nowadays, doctors generally consider the PSA value or the related 1 or 5% - confidence intervals as an important tumor marker of the prostate carcinoma. As the calcium, phosphate and PSA value are feedback controlled by the human body, it is doubtful to consider one-dimensional confidence intervals, thus ignoring correlation.
- Fixing a (nominal) probability of the error of first kind, α , and testing for various hypotheses based on the same data set increases strongly the effective underlying α . For instance, checking 20 hypotheses and fixing $\alpha_{nom} = 5\%$ leads to $\alpha_{eff} \approx 64\%$. In the long run every hypothesis is “accepted”. There exists a Bonferroni correction, but this must be handled with care, too.

5.1 Benford's Law

In the following we focus on numerical data fraud including manipulation and fabrication of figures. It should be stressed that we shall present only a real subset of statistical and related methods. Clustering, classification, Generalized Linear Models or even case-based reasoning truly belong to any anti-fraud tool box. For instance, the last methodology is treated in the context of fraud detection in [46].

Consider a “homogeneous” data set or corpus like revenue and expenditure transactions in business, assurance claims in health care, main economic indicators of an UNO membership country etc. It was *Newcomb* (1881) who first described the phenomena when detecting unexpectedly many usage spots at the digit 1 of some tables of logarithms at hand, [36]. *Benford* (1938) gave the first formal proof, [37], and later on, *Hill* (1995) added more technical details, [38]. Very interesting examples from a broad range of domains can be found in [39].

Simply speaking, the **Benford's Law** states that the distribution of the leading numeral D_1 with range $1, 2, \dots, 9$ of figures from a well defined corpus \mathbf{C} obeying the law is given by $P_C(D_1 = d) = \log(1 + 1/d)$. The formula can be generalized to the first k digits, i.e. $P_C(D_1 = d_1, D_2 = d_2, \dots, D_k = d_k) = \log(1 + 1/(d_1 d_2 \dots d_k))$. Of course, not all data sets are distributed according to the Benford's Law. Sets of identifiers used in database systems or a taxonomy or house numbers are counter examples. Pinkham (1961) proved that for the distribution f_X to be a *Benford* probability distribution it is equivalent to be scale and basis invariant [40]. Let us illustrate the law, and assume a betrayer manipulated a set of bookings by adding faked ones with amounts between Euro 6100 – 6900. A plot of the empirical and Benford's Law signals the forgery, see Fig. 13.

5.2 Data-Model Conforming Tests (DMCTs)

Functional relationships between main business indicators are mostly based on the four arithmetic operations. If we assume that the included variables are superimposed by observational or measurement errors they can be modeled as

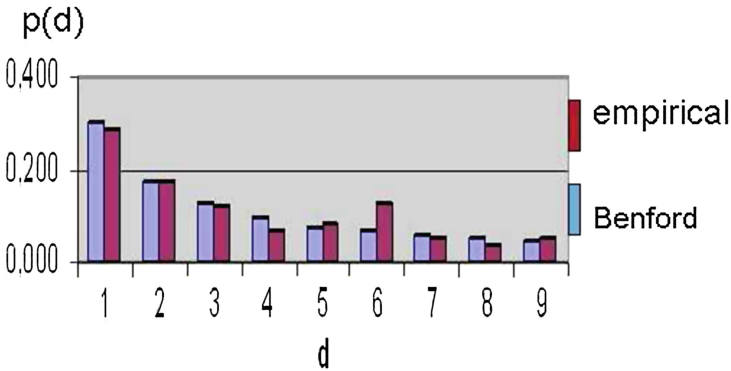


Fig. 13. Manipulated bookings with too many items between Euro 6100 – 6900.

random variables. Consequently, their functional dependency is captured by a (generalized) regression model with errors in the variables, [45]. In the linear case (addition and subtraction) the unobservable (true) values of the variables can be estimated exactly using the **Method of Generalized Least Squares (GLS)**. However, if the variables are linked by multiplication and division a *Taylor Approximation of first kind* developed around the mean vector becomes necessary.

Let $(x, z)_0 \in \mathbf{R}^{p,q}$ be noisy *observation vectors* of the state equation system $x = \xi + u$, and balance equation system $z = \zeta + v = H(\xi) + v$ where $H: \mathbf{R}^p \rightarrow \mathbf{R}^q$. Note, that we use small letters for random vectors. If the balance equations model linear relationships they reduce to the matrix equation $\zeta = H\xi$. For example, in the univariate case we have *Profit = Revenues - Expenditures* $= H\xi = (1, -1) \begin{pmatrix} \text{Revenues} \\ \text{Expenditures} \end{pmatrix}$ representing a linear equation. The fundamental economic relation *Turnover = Quantity \times Price* $= H(\text{Quantity}, \text{Price})$ is of non-linear type.

As proven in [45], $(\xi, \zeta) \in \mathbf{R}^{p,q}$ can be estimated by GLS with maximum precision under the hypothesis of Gaussian noise, a linear system and uncorrelated noise, i.e. $(u, v) \sim N(0, \Sigma_{uv})$. From the computationally point of view the estimation problem reduces to minimizing a quadratic form:

$$\hat{\xi}_{GLS} = \operatorname{argmax}\{(uv)^T \Sigma_{uv}^{-1} \begin{pmatrix} u \\ v \end{pmatrix}\} \quad (1)$$

and

$$\hat{\zeta}_{GLS} = H\hat{\xi}_{GLS}. \quad (2)$$

The variances on the diagonal of the covariance matrix Σ_{uv} are assumed to be completely known due to prior information while the correlations are set to zero. This means that for each variable the observational or measurement error should be known. To some reasonable extent this can be justified by the principle **Minimal Specificity** saying that the covariance matrix Σ_{uv} has a block structure, and all correlations (off-diagonal-elements) vanish. The estimators above have minimal estimation variance according to the **Gauss-Markov Theorem**, [48]. If the relationship is linear but non-normality must be assumed $\hat{\xi}_{GLS}, \hat{\zeta}_{GLS}$ are still best linear unbiased estimators. However, if products or ratios exist as operators, the Gauss-Markov Theorem is only approximately true. In any case the estimates have some convenient characteristics for detecting data-model non-conformity, [45]:

1. $(\hat{\xi}, \hat{\zeta})$ fulfill – except for numerical imprecision – the system of balance equations and
2. $\hat{\Sigma}_{\hat{\xi}} \leq \Sigma_x$ and $\hat{\Sigma}_{\hat{\zeta}} \leq \Sigma_z$ where relational operator “ \leq ” is to be applied to each single component.
3. “Large” deviations between an observational value and its corresponding estimate are a hint to non-conformity and to reject the data-model consistency. This can be statistically tested, [45]
4. Error free variables are not changed.

| c:\qr\modelle\dupont.sht | | | |
|--------------------------------------|---------------------------|-----------------------|---|
| Umsatz 100 ± 5 ? | | Kosten 80 ± 4 ? | Kapital 80 ± 4 ? |
| | Gewinn 30.0 ± 1.5 ? | | Return on investment (%) 40.0 ± 2.0 ? |
| Umsatzrendite (%) 20.0 ± 1.0 ? | | | Kapitalumschlag (%) ? ? |
| | | | |
| | | | |

Fig. 14. Reduced DuPont-System as a spreadsheet.

Let us consider a simple illustrating example. We consider the famous **DuPont-Model** which is presented in Fig. 14. We focus on the (linear) balance equation $Sales = Cost + Profit$. Note that we used German labels for the variables of the model. More formally, $\zeta = \xi_1 + \xi_2$. Let the measurements be imprecise. The observations and the absolute errors of the three quantities are as follows: $Sales(z) = 100 \pm 5$, $Cost(x_1) = 80 \pm 4$ and $Profit(x_2) = 30 \pm 1.5$.

Evidently, the measurements (x_1, x_2, z) do not satisfy the balance equation because $100 \neq 80 + 30$. GLS estimation using the software *Quantor* ([49]), delivers the following consistent estimates and estimation errors: $\hat{\zeta} = 110 \pm 3$, $\hat{\xi}_1 = 85 \pm 3$ and $\hat{\xi}_2 = 25.6 \pm 0.9$. We confirm $110 \approx 85 + 25, 6$. As mentioned above data-model consistent estimates reduce the imprecision of a data set (or leave it unchanged) given levels of 90, 95, 99%-confidence. For instance, $\varepsilon_z = 5 > \hat{\varepsilon}_{\hat{\zeta}} = 3$. Alternative approaches of *GLS* are **Fuzzy Logic** or **MCMC Simulation**. Details can be found in [2].

5.3 Inlier Detection

The *Benford Law* makes clear that a trickster who tries to manipulate numerical data must be careful and skillful doing so because the law imposes logical restrictions on a “human number generator” as we have seen. Generally, further tests on data manipulation and fabrication exist implying more and more restrictions on fudged figures.

Some betrayers fearing to be detected because of generating too large numbers, prefer to do the opposite, i.e. produce “too many” numbers near the average. Such a value is called **Inlier** in Statistics. Roughly speaking, inliers represent a dense cluster of data items around the mean related to a different density.

A log-score approach of inlier detection under the quite strong assumption of independence among the random variables considered is due to *Weir and Murray* (2011) and leads to a “quick and dirty” treatment of the problem, [41]. The procedure is as follows.

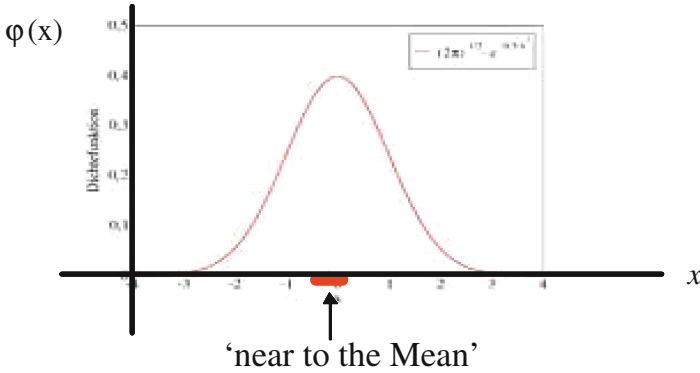


Fig. 15. Potential inlier range under a Gaussian regime.

Procedure Inlier Test

{Weir and Murray (2011)};

Input: Problem size (n,p),
 confidence level (1-alpha),
 data matrix $X_{(n \times p)}$;

Output: Scores s_{zi}^2 , $\ln s_{zi}^2$ for $i=1,2,\dots,n$

begin

Standardize data by $z=(x-my)/\sigma$
 for all p variables and n test objects;

Compute squared score z_i^2 by summing z_{ij}^2
 over all p variables and n objects;

Perform χ^2 test;

Plot \ln -scores $\ln z_i^2$ against $i=1,2,\dots,n$

end.

We illustrate the approach by a fictive case study: Consider the energy consumption (electric power and gas) of Company X. Thus we have $p = 2$ variables of interest. The company runs five factories in each of five districts. Altogether, we have $n = 25$ objects of interest. As the management has suspicion whether or not the factories of a specific district falsify their figures the score test supports assessing the risk of data manipulation.

Figure 16 gives a hint that in district no 3 inliers are present, and that the reported data may be manipulated. While the overall mean of the five districts is $\ln \bar{z}^2 = +0,29$, we observe the mean of district no 3 to be exceptionally small, $\ln z_3^2 = -1,53$. In such cases the management of company X should start trouble shooting for finding out the causality of what has happened.

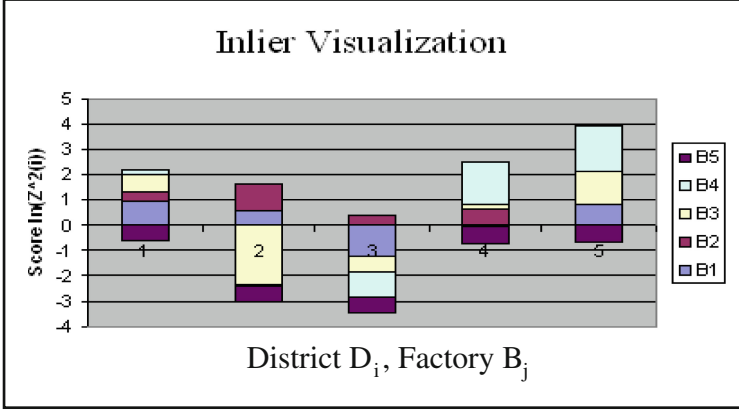


Fig. 16. Hint of inlier generation at all factories of district 3.

5.4 Outlier Tests

Now we turn to the problem of detecting outliers. The motive of tricksters producing outliers either by manipulation or fabrication of data is typical profit making when participating in investment banking, stock market selling or private borrowing.

A naive approach for detecting outliers is given by the “popular” **3-sigma Rule**. Let us assume that the amounts of transactions, say, have a normal distribution, i.e. $X \sim N(\mu, \sigma^2)$. Note, that generally the moments of the Gaussian distribution are unknown and must be estimated from the sample x_1, x_2, \dots, x_n , mostly assuming *identically* and *independently distributed* (i.i.d) observations. The hypothesis of the test is $H_0 : x$ generated by $N(\mu, \sigma^2)$. The alternative hypothesis is $H_1 : x$ not generated by $N(\mu, \sigma^2)$. Under a Gaussian regime the confidence level is $(1 - \alpha) = 0,9973$ for a 3σ -confidence interval. Consequently, the rejection area of the related outlier test given a suspectable observation $x \in \mathbf{R}$ is reject H_0 if

$$|x - \mu|/\sigma > 3. \quad (3)$$

As the unknown parameters must be estimated from the sample which possibly includes outliers the **Masking Effect** is caused, [42]. It leads to the masking of outliers by distorting the estimated mean and standard deviation. The masking effect can be quantified as follows. The 3σ -rule is ineffective to locate outliers with the same sign at a rate $p_I = 1/(1 + \lambda^2)$, [42]. For example, let $\lambda = 3$. It follows $p_I = 10\%$. We visualize the masking effect in Fig. 17.

Evidently, the outlier x_{11} can only be certainly located if the true value of (μ, σ) is known. Even using robust estimation by substituting the standard deviation s by the median (MED) of the absolute deviations around the overall median \tilde{x} as Hampel (1985) proposed, i.e. $MAD(x_\nu)_{\nu=1,2,\dots,n} - \text{MED}(|x_\nu - \tilde{x}|)_{\nu=1,2,\dots,n}$, no convincing improvement of the outlier detection is recognizable.

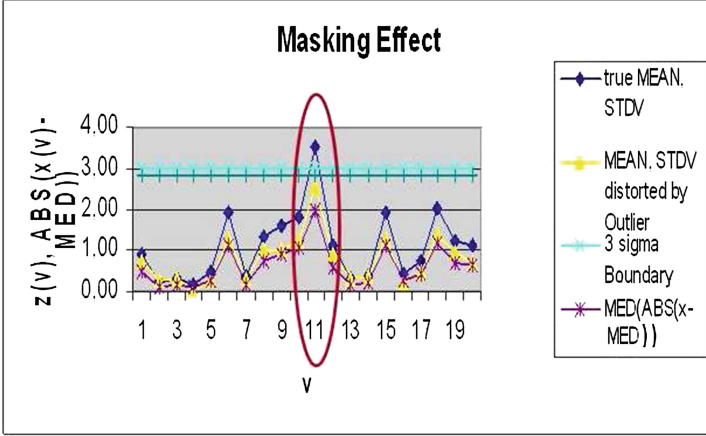


Fig. 17. Masking effect of outlier detection.

In our context, a more efficient outlier location rule is the *c-MAD Rule* proposed by [42]. The idea is to determine the proper $(1 - \alpha_n)$ percentile c of the distribution of $MAD(x_\nu)_{\nu=1,2,\dots,n}$ by Monte-Carlo simulation. Reject $x \in \mathbf{R}$ if

$$|x - \tilde{x}| > c_{n,\alpha_n} MAD(x_\nu)_{\nu=1,2,\dots,n} \quad (4)$$

where $\alpha = \alpha_n = 1 - (1 - \tilde{\alpha})^{1/n}$ and c_{n,α_n} is simulated by solving the inequality

$$P_{n,\alpha_n}(X \notin \{x \in \mathbf{R} \mid |x - \tilde{x}| > c_{n,\alpha_n} MAD(x_\nu)_{\nu=1,2,\dots,n}\}) \geq 1 - \tilde{\alpha} \quad (5)$$

where $\tilde{\alpha} \in (0.5, 1)$ is given. Davies and Gather(1993) showed by example that for a nominal value of $\tilde{\alpha} = 5\%$ the critical values of c_n should be set equal to $c_{20} = 3.02$, $c_{50} = 3.28$, and $c_{100} = 3.47$, [44]. This means that except for very small sample sizes the 3σ -rule is misleading.

A straightforward generalization of outlier detection in multi dimensional data spaces is to use the Mahalanobis distance, cf. [45]. Assume $\mathbf{X} \sim N(\mu, \Sigma)$ with $\mu \in \mathbf{R}^p$ and $\Sigma \in \mathbf{R}^{p \times p}$ known. Then reject $x \in \mathbf{R}^p$ if

$$(x - \mu)^T \Sigma^{-1} (x - \mu) > \chi_{p,1-\alpha}^2. \quad (6)$$

Of course, the same problems as in the univariate case arise:

- unknown parameters (μ, Σ)
- non normality (skewness, mixed distributions)
- outlier robustness.

Alternative approaches to be investigated are given by an *order statistics* or *convex hull confidence regions* approach. The bottleneck here will be the curse of high dimension.

6 Outlook and Perspectives

Data Fraud Detection is a hot topic not only since computers and the Internet dominate our daily life. We conjecture that the relation $\text{fraud rate} \propto 1/\text{ethical attitude}$ is true and there seems some evidence of continuously decreasing ethics in economics, business and society. All important domains like industry, science, public service, press, clinical and pharmaceutical research, health care, religious communities etc. are affected by data fraud. Therefore there exists a great need for improving methodologies and tools for data fraud detection. In science, one step into the right direction of increasing the transparency among the scientific community is by the notification of cross-referencing of ongoing related research. An example for such an authority is the start-up *ResearchGate* at Berlin, Germany.

Furthermore, independent, international scientific data centers are needed where published data must be deposited. The *Principles of Repetability and Reproduceability* are to be obeyed in any case. *Observational and Experimental Selection Bias* should become a taboo. This implies checking the correctness and soundness of experimental designs, collecting data schemes, and applying sound statistical methods. However, such an inspection of data sets of a third party is not a very inspiring task for any researcher!

Finally, we have to admit that there exist vague boundaries between data fraud, fudge, falsification, appraisal, cheat, deception, and scouting. But this should not stop data fraud hunting at all. The scene reminds a bit of the relationship between betrayers and detectives. As the saying goes: “In the long run we get them all!” Or as Di Trocchio put it: “Fraud has been since ever an art. Recently it has become a science.” [6].

References

1. Akkaya, A.D., Tiku, M.L.: Robust estimation and hypothesis testing under short-tailedness and inliers. *Test* **14**(1), 129–150 (2005)
2. Müller, R.M., Lenz, H.-J.: *Business Intelligence*. Springer, Heidelberg (2013)
3. http://de.wissenschaftlichepraxis.wikia.com/wiki/Untersuchungen_zu_Datenfalschung_und_schlechter_Wissenschaft. Accessed: 19 June 2014
4. Gipp, B.: *Citation-based Plagiarism Detection - Applying Citation Pattern Analysis to Identify Currently Non-Machine-Detectable Disguised Plagiarism in Scientific Publications*. Doctoral Dissertation, Technical University Magdeburg (2013)
5. German Press Agency (dpa): DEBEKA Polizei im Haus. *Der Tagesspiegel*, Nr. 22106, July 17, (2014) 15
6. di Trocchio, F.: *Der große Schwindel Betrug und Fälschung in der Wissenschaft*. Campus-Verlag, New York (1994)
7. Sheldrake, R.: *Sieben Experimente, die die Welt verändern*. Scherz Verlag, Berlin (1996)
8. Broad, W.J., Wade, N.: *Betrayers of the Truth*. Oxford Paperback Reference, Oxford (1985)
9. Westfall, R.S.: Newton and the fudge factor. *Sci.* **179**(4079), 751–758 (1973)
10. World System. <http://de.wikipedia.org/wiki/Almagest>. Accessed 18 July (2014)

11. Plotemy. <http://en.wikipedia.org>. Accessed 18 July 2014
12. Almagest. <http://ibiblio.org>. Accessed 18 July 2014
13. Hipparchos. <http://myastrologybook.com>. Accessed 18 July (2014)
14. Galilei, G.: <http://wundervollesrom.com>. Accessed 18 July 2014
15. Galileo's document on heliocentric System. <http://medienwerkstatt-online.de>. Accessed 18 July 2014
16. Galileo's Telescope. <http://www.museum.vic.gov.au/scidiscovery/images/mn006309w150.jpg>. Accessed 18 July 2014
17. Galileo's inclined plane. http://sciencedemonstrations.fas.harvard.edu/icb/icb.do?keyword=k16940&pageid=icb.page80863&pageContentId=icb.pagecontent341734&state=maximize&view=view.do&viewParam_name=indepth.html. Accessed 18 July 2014
18. Godfrey, K.: Newton Portrait. National Portrait Gallery, London (1702)
19. Nature: Austria's most serious report of scientific misconduct in recent memory must be handled properly. (The scandalous behavior of Dr. H. Strasser at MUI, Innsbruck, Austria) *Nature* 454, pp. 917–918, 21 August 2008
20. Ch. Siedenbiedel: Die Libor-Bande. *Frankfurter Allgemeine Sonntagszeitung (FASZ)*, pp. 21–22, no 6, 10 February 2013
21. Tilburg Univ.: Prof. Diederik Stapel suspended. Press release Tilburg University, September 7, (2011). Accessed 17 September 2011
22. Howden, D.: St. Castle: Greece admits deficit figures were fudged to secure Euro entry. *The Independent*, 16 November 2004
23. Gennies, S.: ADAC gibt Manipulationen zu. *Der Tagesspiegel*, vol. 15, no. 21933, 20 January 2014
24. S. Alvarez, J. Huber: Mit dem Zweiten trickst man besser. *Der Tagesspiegel*, vol. 24, no. 22402, 13 July 2014
25. Sanderson, R.: The scandal at the Vatican bank. *Financial Time magazine*, 6 December 2013. <http://www.ft.com/cms/s/2/3029390a-5c68-11e3-931e-00144feabdc0.html>. Accessed 21 July 2014
26. Organ Transplantationen. *Der Spiegel*, pp. 42–44, no 3 (2013)
27. Hartmann, G.: Die Doping-Uni vertuscht ihre Doping-Vergangenheit. *Zeit Online section Sport*, 7 February 2013
28. Ziegler, G.M.: Keine Wahl. In: *Der Tagesspiegel*, Nr. 21816, vol. 31, 21 September 2013
29. Thornton, D., van Capelleveen, G., van Hillegersberg, J., Mueller, R.M.: Outlier-based health insurance fraud detection for u.s. medicaid data. In: *Proceeding of CD on Special Session on Information Systems Security - ISS, ICEIS 2014, Lisboa (2014)*
30. Gipp, B., Meuschke, N., Breitingner, C., Pitman, J., Nürnberger, A.: Web-based demonstration of semantic similarity detection using citation pattern visualization for a cross language plagiarism case. In: *Proceeding of (CD) on Special Session on Information Systems Security - ISS, ICEIS 2014, Lisboa (2014)*
31. *GuttenPlag Wiki* (2011). <http://de.Guttenplag.wikia.com/wiki/GuttenPlag.Wiki>. Accessed 22 July 2014
32. Martinson, B.C., Anderson, M.S., de Vries, R.: Scientists behaving badly. *Nature* 435, pp. 737–738, 9 June 2005. <http://www.vub.ac.be/phd/doctoralschools/lsm/docs/435737a.pdf>. Accessed 20 February 2014
33. John. L.: Seven Shades of Grey. *Psychological Science*, April 2012
34. Ioannidis, J.P.A.: Why most published research findings are false. *PLoS Med.* **2**(8), 0696–0701 (2005)

35. Martin, A.: Five allegations against riken stem-cell researcher in Japan. *Japan Realtime Technology*, 12 March 2014
36. Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. *Am. J. Math.* **4**(1), 39–40 (1881)
37. Benford, F.: The law of anomalous numbers. *Proc. Am. Phil. Soc.* **78**, 551–572 (1938)
38. Hill, T.P.: A statistical derivation of the significant-digit law. *Stat. Sci.* **10**, 354–363 (1995)
39. Nigrini, M.J.: *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley, New York (2012)
40. Pinkham, R.S.: On the distribution of first significant digits. *Ann. Math.Stat.* **32**(4), 1223–1230 (1961)
41. Weir, C., Murray, G.: Fraud in clinical trials detecting it and preventing it. *Signif.* **8**(4), 164–168 (2011)
42. Davies, L., Gather, U.: Robust statistics. In: Gentle, J., Härdle, W., Mori, Y. (eds.) *Handbook of Computational Statistics Concepts and Methods*, pp. 655–695. Springer, Heidelberg (2004)
43. Hampel, F.: The breakdown points of the mean combined with some rejection rules. *Technometrics* **27**, 95–107 (1985)
44. Davies, L., Gather, U.: The identification of multiple outliers. *J. Am. Stat. Assoc.* **88**, 782–801 (1993)
45. Lenz, H.-J., Röel, E.: Statistical quality control of data. In: Gritzmann, P., Hettich, R., Host, R., Sachs, E. (eds.) *Operations Research 1991*, pp. 341–346. Springer, Heidelberg (1991)
46. Wheeler, R., Aitken, S.: Multiple algorithms for fraud detection. *Knowl. Based Syst.* **13**(2–3), 93–99 (2000)
47. Windolf, P.: Korruption, betrug und ‘corporate governance’ in den USA - anmerkungen zu enron. *Leviathan* **31**(2), 185–218 (2003)
48. Anderson, T.W.: *An Introduction to Multivariate Statistical Analysis*. Wiley, New York (1958)
49. Schmid, B.: Bilanzmodelle. Simulationsverfahren zur Verarbeitung unscharfer Teilinformationen. ORL-Bericht No.40, ORL Institut, Universität Zürich (1979)
50. Mosler, K., Lange, T., Bazovkin, P.: Computing zonoid trimmed regions in dimension $d > 2$. *Comput. Stat. Data Anal.* **53**, pp. 2500–2510 (2009)

Enterprise Information Systems

16th International Conference, ICEIS 2014, Lisbon,
Portugal, April 27-30, 2014, Revised Selected Papers
Cordeiro, J.; Hammoudi, S.; Maciaszek, L.A.; Camp, O.;
Filipe, J. (Eds.)

2015, XVII, 490 p. 178 illus., Softcover

ISBN: 978-3-319-22347-6