Syntactic and Semantic Bias Detection and Countermeasures

Roman Englert[1,2(✉)] and Jörg Muschiol[1]

[1] FOM University of Applied Sciences,
Herkulesstraße 32, 45127 Essen, Germany
roman.englert@fom-net.de
[2] Faculty III, New Media and Information Systems, Siegen University,
Kohlbettstraße 15, 57072 Siegen, Germany

Abstract. Applied Artificial Intelligence (AAI) and, especially Machine Learning (ML), both had recently a breakthrough with high-performant hardware for Deep Learning [1]. Additionally, big companies like Huawei and Google are adapting their product philosophy to AAI and ML [2–4]. Using ML-based systems require always a training data set to achieve a usable, i.e. trained, AAI system. The quality of the training data set determines the quality of the predictions. One important quality factor is that the training data are unbiased. Bias may lead in the worst case to incorrect and unusable predictions. This paper investigates the most important types of bias, namely syntactic and semantic bias. Countermeasures and methods to detect these biases are provided to diminish the deficiencies.

Keywords: Bias detection · Training samples · Multivariate regression · Root-out-bias

1 Introduction

The term bias has several meanings: in AI and Machine Learning (ML) "any preference for one hypothesis over another, beyond mere consistency with the examples, is called a bias" [5]. In other words a (declarative) bias helps to understand how prior knowledge can be used to identify the hypothesis space within which to search. The bias is independent of the applied ML technique, i.e. probability theory or (inductive) logic programming. In psychology is the omission bias "the tendency to judge harmful actions as worse, or less moral than equally harmful omissions (inactions) because actions are more obvious than inactions" [6], and in mathematics is a bias a systematic error.

An applied understanding of bias is prejudice by morally incomplete data, where the application is in training ML algorithms/models with data sets. As an example serves an American computer scientist discovering that Google's facial recognition Software only spotted his face, if he wore a white mask (see Fig. 1) [7]. Since ML models provide predictions based on the used training data sets, potential biases need to be recognized and defined before training data sets are being generated.

V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12140, pp. 629–638, 2020.
https://doi.org/10.1007/978-3-030-50423-6_47

Fig. 1. An American computer scientist, found his computer system recognized the white mask, but not his face.

Bias detection requires an investigation of the training sample. In the case of the American computer scientist, the bias is called semantic bias, since a feature is missing in the training data. If the training sample contains mathematical computable biases like features are dependent on each other, or heteroscedasticity [8], then the bias is called syntactic. Both types of bias require dedicated methods to achieve a bias mitigation, the former the root-out-bias method with an interrogation by a human expert [9], and the latter a pre-processing of the training data [10–12]. For the mitigation of bias we focus on data transformation techniques.

The paper is structured as follows: The importance of AAI and training data is described in Sect. 2. Then, the computation of syntactic bias including the state-of-the art of research is described in Sect. 3. Subsequently, Sect. 4 contains an example for training data that became insignificant after bias inspection. And Sect. 5 contains the detection of missing features with the root-out-bias method and an example based on an interrogation by a human expert. Section 6 concludes the paper with a summary and an outlook discussing further research for the automatization of syntactic and semantic bias detection.

2 AAI Strategies of Google and Huawei and the Importance of Proper Training Samples

After a long period of research, AI reached a maturity level and had 2007 the breakthrough with high-performant Deep Learning chips developed by Huawei and Google [1]. These two leading companies announced to focus their R&D completely onto AI [2–4]. Their AI strategies are described in the following, since they influence the applied data science world and, thus, highlight the importance of proper training data. An example for a famous improper training sample, the Wooldridge data set "affairs" [13], is shown in the subsequent section.

Huawei started focusing on AI already in the eighties with research on AI algorithms for classification (e.g. regression), social analytics (e.g., PageRank), dimensionality reduction (e.g., KPCA), anomaly detection (e.g., local outlier factor), and clustering of samples (e.g., K-means, GMM), to name a few important ones. Huawei's AI strategy is based on AI research with the following focus fields [4]:

- Image processing and interpretation
- Natural language understanding
- Decision making and inferences from knowledge domains

Accompanying goals are an optimization of required data, less computational effort and less energy consumption. Additionally, Machine Learning is aimed to be secure and trustworthy, and should be processed autonomously and fully automated. In order to reach this comprehensive AI goal Huawei implements four measures:

1. A full-stack AI portfolio consisting of distributed cloud, devices, algorithms and applications for multi-purposes.
2. The development of a talent hub through the cooperation with universities and industry.
3. Huawei's portfolio is completely based on AI.
4. All processes are based on AI and an efficiency increase is expected.

An example for a widely known process is email spam filtering. This strategy will influence research at universities and development.

Google has a different AI strategy [3, 14]: As hitherto they had a focus onto the search engine and data collection based on services like email, Google Scholar and Maps/Earth. Today Google offers cloud-based ML applications [3]:

- Contact Center AI: A trainable customer support system for the automated communication.
- Document Understanding AI: ML-based understanding of documents using text extraction and information tagging.
- Cloud Talent Solution: Matching of job offers and applicants.
- Recommendations AI: Personalized product recommendations with real-time adaptations to customer behavior.

Thus, the former focus fields are becoming less important, and data are in future collected through cloud-based ML applications. Google focuses on research in the areas of Deep Learning (Neural Networks with several hidden layers), document analysis, Pattern Recognition, feature (text) extraction, recommendations, and (probabilistic) matching. Adaptive AI systems require huge training samples without bias. The aforementioned AI applications reason the importance for proper unbiased training samples.

3 Syntactic Bias and Its Mitigation

The preprocessing of training samples to mitigate bias is described in varies papers [10–12]. Some frequent occurring anomalies are discussed in the following (the mathematical representation is based on [8]):

1. Normal distribution: If features are normally-distributed and the sample size is big enough, i.e., the central limit theorem holds, then varies tests like the t-test can be applied. The t-test determines, whether the average of a random sample deviates more than a given value p from the population mean.
2. Significance: Features (variables) of training samples that are not significant according to, e.g., the 5% level, can be dropped. Assume that the considered feature of the training sample is normally-distributed. Then, the f- and the t-test can be combined used to check the significance of the feature: Assume p is 5%, then the variable can be dropped, if the p-level of the t-test is less equal 0.05, and the f-test shows that the regressors coefficients are equal (to zero). The latter one investigates the joint significance of the features.
3. Dummy traps: This effect holds for regression, when the model suffers from (multi) collinearity. In this case can one variable be (linearly) predicted of the others. As example consider the variables x, y, and z, then these variables are collinear, iff

$$x = a \cdot y + b \cdot z. \tag{1}$$

   with a and b are real-valued vectors. In this case, the estimate of the impact of variable x onto another dependent variable is less precise compared to the situation without collinearity, and the variable x should be dropped.
4. Independent and identical distributed: This property for distributions of variables is important for the central limit theorem, stating that the probability distribution of the sum of independent and identical distributed variables (with finite variance) approximates the normal distribution. As a consequence, e.g., the significance of features can be investigated.

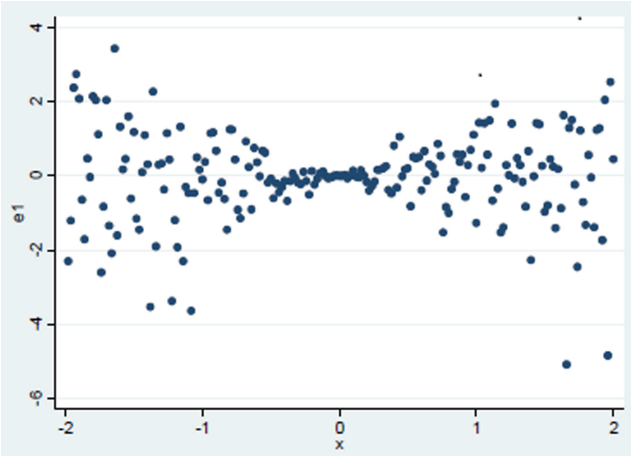

Fig. 2. An example for heteroscedasticity [15].

5. Heteroscedasticity: In this case the dispersion of sub-populations differs (see Fig. 2), and statistical hypothesis tests for their significance can become invalid. Heteroscedasticity can be detected by applying the f-test, which identifies how good a model fits to the training sample, since the f-test compares the variance of two sub-samples. Especially for non-linear models may occur severe impacts, where models can become inconsistent. As an example consider the speed measurement of a moving object. The measurement close to the object is more precise than if the object has a greater distance. Thus, the measurement data are expected to contain heteroscedasticity, which is a general challenge of measurements [8]. Heteroscedasticity can be mitigated using various approaches. An overview of detecting heteroscedasticity and mitigating its effects is provided in [22]. The mitigation is based on an analysis and transformation of the residuals that cause the heteroscedasticity effect.

The above discussed anomalies can also be mitigated by applying mappings of random samples to new distributions with constraints [11]. In the following the mitigation of bias using an inspection tool is described. As exemplary tool, the What-if tool from Google is used [16].

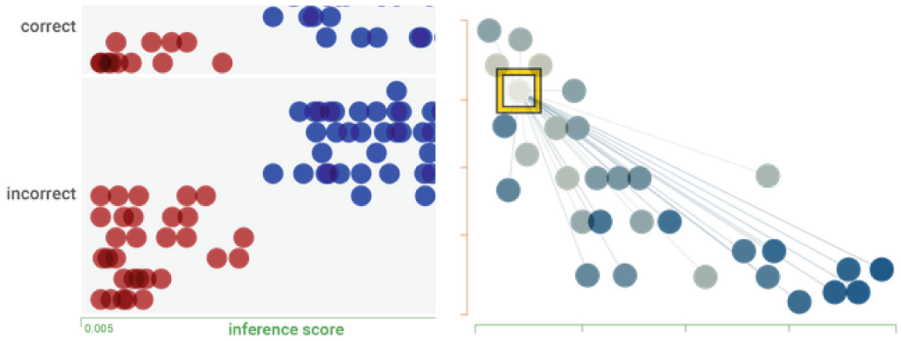

Fig. 3. What-if tool: visualization of inference results and arrangement of data points by similarity [16].

The What-if tool supports model inspection with the visualization of inference results with correct/incorrect data points (false/positives) and to compare two models (Fig. 3, left). Furthermore, data points can be arranged by similarity, whereat the user can create distance features and apply them to the model for inspection (Fig. 3, right).

Fig. 4. Re-weighing of data points and comparison of counterfactuals to data points [16].

The mitigation of bias is supported by re-weighing of features (Fig. 4, left). However, this may require some iterations and trials, and may become a tedious task. Analogously, the possibility to compare data points and their counterfactuals to gain insight into the portion of data points to model, resp. the statistical distribution of features. Nevertheless, this manual inspection may be tedious and requires expert knowledge and experience, but automated tools like Fairness 360° [18] require that an expert chooses a re-weighing measure among more than 70 choices and inspects the result of its application to the training sample. Bias mitigation and syntactic bias detection, both, require a tool and a human expert.

4 An Example for Training Data that Became Insignificant After Bias Inspection

Social contacts are an important component of an individual's life and many adults follow the wish to marry and start a family [13]. However, it is still an unsolved question whether the human being suits monogamy, since the "happily ever after" feeling does not always last and a partner may engage in an extramarital affair. For the evaluation the probability is investigated of having an affair based on data set "affairs".

In 1969 and 1974 data from two magazine surveys have been collected about men and women time spent with beaus. From these surveys 601 samples have been selected based on the criteria that the interrogated people were employed and married for the first time (Fig. 5).

ID	male (=1)	age	years married	kids (1=yes)	relig (1-5)	educ	occup	rate marriage	# affairs last year	vryhap	hapavg	avgmarr	vry reli	smereli	slightreli	had affair(1= yes)	unhappy	notrel
4	1	37	10	0	3	18	7	4	0	0	1	0	0	0	1	0	0	0
5	0	27	4	0	4	14	6	4	0	0	1	0	0	1	0	0	0	0
6	1	27	1,5	0	3	18	4	4	3	0	1	0	0	0	1	1	0	0
11	0	32	15	1	1	12	1	4	0	0	1	0	0	0	0	0	0	0
12	0	27	4	1	3	17	1	5	3	1	0	0	0	0	1	1	0	0
16	1	57	15	1	5	18	6	5	0	1	0	0	1	0	0	0	0	0
23	1	22	0,75	0	2	17	6	3	0	0	0	1	0	0	0	0	0	1
29	0	32	1,5	0	2	17	5	5	0	1	0	0	0	0	0	0	0	1
43	1	37	15	1	5	18	6	2	7	0	0	0	1	0	0	1	1	0
44	0	22	0,75	0	2	12	1	3	0	0	0	1	0	0	0	0	0	1
45	1	57	15	1	2	14	4	4	0	0	1	0	0	0	0	0	0	1
47	0	32	15	1	4	16	1	2	0	0	0	0	0	0	1	0	1	0
49	1	22	1,5	0	4	14	4	5	0	1	0	0	0	1	0	0	0	0
50	1	37	15	1	2	20	7	2	0	0	0	0	0	0	0	0	1	1
53	0	32	10	1	3	17	5	2	12	0	0	0	0	0	1	1	1	0
55	1	27	4	1	4	18	6	4	0	0	1	0	0	1	0	0	0	0
64	1	47	15	1	5	17	6	4	0	0	1	0	1	0	0	0	0	0
67	1	22	0,125	0	4	16	5	5	1	1	0	0	0	1	0	1	0	0
79	0	22	1,5	1	2	14	1	5	1	1	0	0	0	0	0	1	0	1
80	0	22	1,5	0	2	17	5	4	0	0	1	0	0	0	0	0	0	1
86	0	27	4	0	4	14	5	4	0	0	1	0	0	0	1	0	0	0
93	0	37	15	1	1	17	5	5	0	1	0	0	0	0	0	0	0	0
108	0	37	15	1	2	18	4	3	0	0	0	1	0	0	0	0	0	1
114	0	22	0,75	0	3	16	5	4	0	0	1	0	0	0	1	0	0	0
115	0	22	1,5	0	2	16	5	5	0	1	0	0	0	0	0	0	0	1
116	0	27	10	1	2	14	1	5	0	1	0	0	0	0	0	0	0	1
122	1	37	15	1	4	14	5	2	12	0	0	0	0	1	0	1	1	0
123	0	22	1,5	0	2	16	5	5	0	1	0	0	0	0	0	0	0	1
126	0	22	1,5	0	2	14	3	4	7	0	1	0	0	0	0	1	0	1
127	0	22	1,5	0	2	16	5	5	0	1	0	0	0	0	0	0	0	1
129	0	27	10	1	4	16	5	4	0	0	1	0	0	1	0	0	0	0
133	1	37	15	1	2	18	6	4	2	0	1	0	0	0	0	1	0	1
134	0	32	10	1	3	14	1	5	0	1	0	0	0	0	1	0	0	0
137	1	37	4	1	2	20	6	4	0	0	1	0	0	0	0	0	0	1
138	0	32	15	1	4	12	3	2	3	0	0	0	0	1	0	1	1	0
139	0	22	1,5	0	2	18	5	5	0	1	0	0	0	0	0	0	0	1
147	0	27	7	0	4	16	1	5	0	1	0	0	0	1	0	0	0	0

Fig. 5. 601 samples from Wooldridge's data set "affairs" (excerpt). Dummy traps are highlighted red and the regressand yellow. (Color figure online)

4.1 Data Preprocessing

The data set is provided as an EXCEL file that consists of 601 rows containing the answers of the interrogated persons and nineteen columns with features, i.e., nineteen candidates for variables. Since the probability of having an affair has to be investigated, the column "id" that identifies an individual will be ignored for the regression computation, and column "affair" (=1, if had an affair) is taken as regressand (the dependent variable that is explained), the dependent random variable for the regression model. First, the data are sharpened by dropping those variables that are not significant according to the 5% level, i.e., p-level of t-test ≤ 0.05 (see also Sect. 3).

Furthermore, the data set contains twice a dummy trap, i.e., the model suffers from collinearity (cf. Fig. 5, red highlighted). First, columns twelve till fifteen contain a zero or one ("vryhap", "hapavg", "avgmarr", "unhap") and are perfect collinear, i.e., each row sums up to one for these four (binary) variables. Hence, the column "unhap" is left out and the interpretation for the remaining three variables is "relative to having an unhappy marriage". Analogous, columns sixteen till nineteen contain a zero or one ("vryrel", "smerel", "slghtrel", "notrel") and are perfect collinear, i.e., each row sums up to one for these four variables. Hence, the column "notrel" is left out and the interpretation for the remaining three variables is "relative to being not religious".

After computing the regression including the t-test (significance test of variables), two variables are determined with a p-value ≤ 0.05, and thus, these variables are not significant and dropped: "kids" and "naffairs". The check for heteroscedasticity was negative. To summarize, thirteen variables remain as regressors (independent variables): "male", "age", "yrsmarr", "relig", "educ", "occup", "ratemarr", "vryhap", "hapavg", "avgmarr", "vryrel", "smerel" and "slghtrel" (cf. Fig. 5).

4.2 Interpretation of the Parameter Estimates of the Significant Variables in the Model

The parameter estimates lead to the following model for the regressand

$$\begin{aligned}\Pr(\hat{y}=1) &= 0.8342 + 0.0511 * male - 0.0073 * age + 0.0181 * yrsmarr \\ &- 0.1856 * relig + 0.0032 * educ + 0.0045 * occup + 0.0102 * ratemarr \\ &- 0.3541 * vryhap - 0.2698 * hapavg - 0.2078 * avgmarr + 0.4884 * vryrel \\ &+ 0.2879 * smerel + 0.2590 * slghtrel\end{aligned} \tag{2}$$

The regressand "affair" is used to estimate the linear probability model. Note, the intercept is 0.8342 (expected mean value, if all variables are 0, however "age" cannot be 0). First, the variables "age", "relig", "vryhap", hapavg", and "avgmarr" have negative effects on the probability of a person to engage in an affair. The marginal effect of the variable "age" is, that when a person turns one year older, it reduces the probability of having an affair by 0.0073. The more religious a person is ("relig": 1…5 (high)), the probability to have an affair will be reduced by 0.1856 times "relig". The dummy variables "vryhap", "hapavg" and "avgmarr" each have a negative effect on the likelihood of a person having a paramour in relation to an unhappy marriage, namely:

−0.3541, −0.2698, −0.2078. Second, the five variables “male”, “yrsmarr”, “educ”, “occup” and “ratemarr” have a small positive coefficient (<0.06): The average effect on the likelihood to have an affair is for a male with a factor of 0.0511 higher than for a woman (“male”). Third, the dummy variables “vryrel”, “smerel” and “slghtrel” have a relatively large positive contribution (0.4884, 0.2879, 0.2590) on the likelihood to have an affair (in relation to being not religious).

The R-squared value is low with 0,1261, meaning that the data explain poorly the regressand. Hence, a RESET test [18] with quadratic and cubic order is recommended in order to test for a more precise fit.

5 Semantic Bias and the Root-Out-Bias Method

Semantic bias is in contrast to syntactic bias not computable and may have different causes that are difficult to detect. As an example consider the American computer scientist who test a face recognition Software and found his computer system recognized the white mask, but not his face (Fig. 1, Sect. 1). The training sample for the utilized Neural Network missed at least one feature, and thus, is biased. The omitted and missing feature is ambiguous, since it might be for instance “skin type” describing characteristics of the skin, or the feature “culture” depicting the cultural background, assuming that the cultural background influences the appearance of someone. This type of bias is called semantic bias.

Another example is the Chabot Tay from Microsoft [20, 21]: The goal with Tay was to perform research and to gain deeper insights into conversational understanding. Tay was automatically trained by its conversations with unknown persons via the Internet. Unfortunately, Tay was trained with biased statements according to gender, and additionally, to racists statements. Microsoft had to take off Tay immediately. This demonstrates that speech could be biased, and must be inspected before presenting the statements to the underlying Neural Network.

A method to identify semantic bias is the so-called “root-out-bias” [9]: Since semantic bias is not computable, a human expert must be involved. The root-out-bias consists of two steps: The first step is to openly question what preconceptions could currently exist in a domain that is aimed to be modeled. This important step requires experience, and must be done by experts. As outcome potential biases are identified, and then, in step two requirements to the training data set must be defined.

In order to illustrate the root-out-bias method it is applied to the Wooldridge data set from Sect. 4. The data set contains 16 features to explain the probability that someone has an extramarital affair. One feature is the sex of an interrogated person, where is assumed that an affair only takes place between different sexes [13]. This assumption is outdated, since 2015 the same-gender marriage was invented in the US [19]. Thus, nowadays the data set is semantically-biased and this assumption should be revised. The new requirement for this data set can be that people are interrogated despite the gender of their marriage partner.

This example demonstrates that semantic bias has sweeping facets and is not only culturally-dependent (also law-dependent), and may change in societies and cultures by time.

6 Summary and Outlook

This paper analyses syntactic and semantic biases within training samples. Syntactic bias can be detected by computation. But the mitigation requires a human expert who can be supported by tools for data inspection and visualization. In contrast, semantic bias cannot be computed and must be detected by a structured procedure, e.g., root-out-bias method by an experienced human interrogator. Varies examples demonstrate the importance of proper training samples to avoid bias. Biased training samples can lead to social and morally unacceptable AI systems that need to be taken off.

As an outlook the investigation for the automatization of the detection and mitigation of syntactic and semantic bias is aimed. For syntactic bias provides the mathematical modeling of biases the "parts" of the training sample that needs to be improved. And semantic bias may be analyzed using semantic networks that brings knowledge items semantically into relation to each other.

References

1. Groth, O., Nitzberg, M.: Solomon's Code. Humanity in a World of Thinking Machines. Pegasus Books (2018)
2. Artificial Intelligence Technology Scan, Teqmine Technology Analysts. https://teqmine.com/google-ai-strategy/. Accessed 28 Oct 2019
3. Google's AI-based Cloud Services. https://cloud.google.com/solutions/ai/?hl=de. Accessed 28 Oct 2019
4. Huawei Releases AI Strategy und Full-Stack, All-Scenario AI Portfolio. https://www.huawei.com/en/press-events/news/2018/10/Huawei-HC-2018-Eric-Xu-AI. Accessed 28 Oct 2019
5. Russell, S., Norvig, P.: AI a Modern Approach. Prentice-Hall, Upper Saddle River (1995)
6. Psychology omission bias definition. https://en.wikipedia.org/wiki/Omission_bias. Accessed 28 Oct 2019
7. Bias in the face recognition Software of Google. https://www.bbc.com/news/technology-45561955. Accessed 28 Oct 2019
8. Casella, G., Berger, R.: Statistical Inference. Duxbury Advanced Series, 2nd edn. Thomson Learning Inc., Boston (2002)
9. Root-out-bias method. https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it. Accessed 28 Oct 2019
10. Kamiran, F., Calder, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8
11. Calmon, F, Wei, D., Vinzamuri, B., Ramamurthy, K., Varshney, K.: Optimized preprocessing for discrimination prevention. In: Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems, vol. 30, pp. 3992–4001 (2017)
12. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. international conference on machine learning. In: Proceedings of Machine Learning Research, pp. 325–333 (2013)
13. Fair, R.: A theory of extramarital affairs. J. Polit. Econ. **86**(1), 45–62 (1978)
14. Google's AI. https://ai.google. Accessed 10 Nov 2019

15. Richard, W.: Heteroskedasticity, University of Notre Dame. https://www3.nd.edu/~rwilliam. Accessed 11 Nov 2019
16. Google's What-If-Tool. https://pair-code.github.io/what-if-tool. Accessed 11 Nov 2019
17. Fairness 360°. https://aif360.mybluemix.net. Accessed 11 Nov 2019
18. Sapra, S.: A regression error specification test (RESET) for generalized linear models. Econ. Bull. **3**(1), 1–6 (2005)
19. Same-gender marriage US. https://www.nytimes.com/2015/06/27/us/supreme-court-same-sex-marriage.html. Accessed 16 Nov 2019
20. Microsoft is deleting its AI chatbot's incredibly racist tweets. https://img.sauf.ca/pictures/2016-03-24/d360716e3199095063ebd4749b78fc4c.pdf. Accessed 16 Nov 2019
21. Davis, E.: AI amusements: the tragic tale of tay the chatbot. ACM AI Matters **2**(4), 20–24 (2016)
22. Rosopa, P., Schaffer, M., Schroeder, A.: Managing heteroscedasticity in general linear models. Psychol. Methods **18**(3), 335–351 (2013)