

# A Preliminary Investigation into Predictive Models for Adverse Drug Events

**Jesse Davis**

KU Leuven  
jesse.davis@cs.kuleuven.be

**Vitor Santos Costa**

Universidade do Porto  
vsc@dcc.fc.up.pt

**Peggy Peissig and Michael Caldwell**

Marshfield Clinic  
{peissig,peggy,caldwell.michael}@marshfieldclinic.org

**David Page**

University of Wisconsin - Madison  
page@biostat.wisc.edu

## Abstract

Adverse drug events are a leading cause of danger and cost in health care. We could reduce both the danger and the cost if we had accurate models to predict, at prescription time for each drug, which patients are most at risk for known adverse reactions to that drug, such as myocardial infarction (MI, or “heart attack”) if given a Cox2 inhibitor, angioedema if given an ACE inhibitor, or bleeding if given an anticoagulant such as Warfarin. We address this task for the specific case of Cox2 inhibitors, a type of non-steroidal anti-inflammatory drug (NSAID) or pain reliever that is easier on the gastrointestinal system than most NSAIDs. Because of the MI adverse drug reaction, some but not all very effective Cox2 inhibitors were removed from the market. Specifically, we use machine learning to predict which patients on a Cox2 inhibitor would suffer an MI. An important issue for machine learning is that we do not know which of these patients might have suffered an MI even without the drug. To begin to make some headway on this important problem, we compare our predictive model for MI for patients on Cox2 inhibitors against a more general model for predicting MI among a broader population not on Cox2 inhibitors.

## Introduction

An electronic medical record (EMR) or electronic health record (EHR) is a relational database that stores a patient’s clinical history: disease diagnoses, procedures, prescriptions, vitals and lab results, etc. EMRs contain a wealth of information, and using techniques such as machine learning and data mining to analyze them offers the potential to discover interesting medical insights. For example, it is possible to use an EMR to build models to address questions such as predicting which patients are most at risk for having an adverse response to a certain drug or predicting the efficacy of a drug for a given individual.

The automated analysis of EMR data poses many challenges, including:

**Multiple relations.** Each type of data (e.g., drug prescription information, lab test results, etc.) is stored in a different table of a database. Traditionally, machine learning algorithms assume that data are stored in a single table.

**Representation of uncertainty.** It is necessary to model the non-deterministic relationships between a patient’s past and future health status.

**Noisy data.** The data are inherently noisy. For example, a disease code may be recorded for billing purposes or because a patient had the disease in the past and the physician found it relevant for the current consultation. Or lab test results may vary due to lab conditions and personnel.

**Incomplete data.** Important information such as the use of over-the-counter drugs may not appear in the clinical history.

**Varying amounts of data.** The amount of data about each patient may vary dramatically. For example, patients switch doctors and clinics over time, so a patient’s entire clinical history is unlikely to reside in one database.

**Longitudinal data.** The time of diagnosis or drug prescription is often very important.

Statistical relational learning (SRL) (Getoor and Taskar 2007) is a relatively new area of artificial intelligence research that is particularly suited to analyzing EMRs. The goal of SRL is develop formalisms that combine the benefits of relational representations, such as relational databases or first-order logic, with those of probabilistic, graphical models for handling uncertainty. SRL is particularly adept at simultaneously modeling the relational structure and uncertain nature of EMRs.

In this paper, we will look at applying SAYU (Davis et al. 2005), a SRL formalism that combines rule learning with probabilistic inference, to the specific task of predicting adverse drug reactions (ADRs) from EMR data. Building predictive models for ADRs is interesting because an accurate model has the potential to be actionable. However, in order to be actionable, one important question is whether the model is predicting susceptibility to the ADR or simply that a patient is at risk for the outcome, regardless of whether the drug (or medicine) is taken. In this paper, we present a preliminary investigation of this question using real-world clinical data. We focus on one ADR related to selective Cox2 inhibitors (Cox2ib), such as Vioxx<sup>TM</sup>.

Our results show significant differences in transferring learned models for myocardial infarction (MI) between two populations, one that was exposed to Cox2 inhibitors and

a control population, and in general suggest differential behavior between the two populations with higher accuracy in models for populations exposed to Cox2 inhibitors.

## Background

In this section, we review SAYU (Davis et al. 2005), an SRL learning system. SAYU uses first-order definite clauses, which can capture relational information, to define (binary) features. These features then become nodes in a Bayesian network.

### First-order Logic

SAYU defines features using the non-recursive Datalog subset of first-order logic.<sup>1</sup> The alphabet consists of three types of symbols: constants, variables, and predicates. *Constants* (e.g., the drug name `Proprano1o1`), which start with an upper case letter, denote specific objects in the domain. *Variable* symbols (e.g., `disease`), denoted by lower case letters, range over objects in the domain. *Predicate* symbols  $P/n$ , where  $n$  refers to the arity of the predicate and  $n \geq 0$ , represent relations among objects. An *atom* is of the form  $P(t_1, \dots, t_n)$  where each  $t_i$  is a constant or variable. A *literal* is an atom or its negation. A *clause* is a disjunction over a finite set of literals. A *definite clause* is a clause that contains exactly one positive literal. Definite clauses are often written as an implication  $B \implies H$ , where  $B$  is a conjunction of literals called the body and  $H$  is a single literal called the head.

### Tree-Augmented Naive Bayes

A Bayesian network compactly represents the joint probability distribution over a set of random variables  $X = \{X_1, \dots, X_n\}$ . A Bayesian network is a directed, acyclic graph that contains a node for each random variable  $X_i \in X$ . For each variable (node) in the graph, the Bayesian network has a conditional probability table  $\theta_{X_i|Parents(X_i)}$  giving the probability distribution over the values that variable can take for each possible setting of its parents. A Bayesian network encodes the following probability distribution:

$$P_B(X_1, \dots, X_n) = \prod_{i=1}^{i=n} P(X_i|Parents(X_i)) \quad (1)$$

Tree augmented naïve Bayes (TAN) (Friedman, Geiger, and Goldszmidt 1997) is a Bayesian network based classifier. Given a set of attributes  $A_1, \dots, A_n$  and a class variable  $C$ , the TAN learning algorithm starts with basic Bayes net structure of naïve Bayes, where each attribute has exactly one parent: the class node. TAN further permits each attribute to have at most one other parent. This allows the model to capture a limited set of dependencies between attributes. A TAN model can be learned in polynomial time with a guarantee that the learned model maximizes the log likelihood of the data given the class node.

<sup>1</sup>This subset of first-order logic with a closed-world assumption is equivalent to relational algebra/calculus.

## SAYU

SAYU, like VISTA (Davis et al. 2007), uses definite clauses to define features for the statistical model. Each definite clause becomes a binary feature in the underlying statistical model. The feature receives a value of one for an example if the data about the example satisfies (i.e., proves) the clause and it receives a value of zero otherwise.

SAYU starts by learning a model  $M$  over an empty feature set  $FS$ . This corresponds to a model that predicts the prior probability of the target predicate. Then it repeatedly searches for new features for a number of iterations. In each iteration, SAYU first selects a random seed example and then performs a general-to-specific, breadth-first search through the space of candidate clauses. To guide the search process, it constructs the *bottom clause* by finding all facts that are relevant to the seed example (Muggleton 1995). SAYU constructs a rule containing just the target attribute, such as `ADR(pid)`, on the right-hand side of the implication. This means that the feature matches all examples. It creates candidate features by adding literals that appear in the bottom clause to the left-hand side of the rule, which makes the feature more specific (i.e., it matches fewer examples). Restricting the candidate literals to those that appear in the bottom clause helps limit the search space while guaranteeing that each generated refinement matches at least one example.

SAYU converts each candidate clause into a feature,  $f$ , and evaluates  $f$  by learning a new model (e.g., the structure of a Bayesian network) that incorporates  $f$ . In principle, any structure learner could be used, but SAYU typically uses a tree-augmented Naive Bayes model (Friedman, Geiger, and Goldszmidt 1997). SAYU evaluates each candidate  $f$  by comparing the generalization ability of the current model  $FS$  versus a model learned over a feature set extended with  $f$ . SAYU does this by calculating the area under the precision-recall curve (AUC-PR) on a tuning set (Davis and Goadrich 2006). AUC-PR is used because relational domains typically have many more negative examples than positive examples, and the AUC-PR ignores the potentially large number of true negative examples.<sup>2</sup> In each iteration, SAYU adds the feature  $f'$  to  $FS$  that results in the largest improvement in the score of the model. In order to be included in the model,  $f'$  must improve the score by a certain percentage-based threshold. This helps control overfitting by pruning relative weak features that only improve the model score slightly. If no feature improves the model's score, then it simply proceeds to the next iteration.

## Data

Our data comes from Marshfield Clinic, an organization of hospitals and clinics in northern Wisconsin. This organization has been using electronic medical records since 1985 and has electronic data back to the early 1960's. We have received institutional review board approval to undertake these studies.

<sup>2</sup>In principle, SAYU can use any evaluation metric to evaluate the quality of the model including (conditional) likelihood, accuracy, ROC analysis, etc.

We included information from four separate relational tables: lab test results (e.g., cholesterol levels), medications taken (both prescription and non-prescription), disease diagnoses, and observations (e.g., height, weight and blood pressure). We only consider patient data up to one week before that patient’s first prescription of the event under consideration. This ensures that we are building predictive models only from data generated before the event occurs. Table 1 provides the characteristics of each patient population that we will use in our case study. It reports the number of unique medication codes, diagnosis codes and observation values that occur in these tables. The table size rows lists the number of facts (i.e., rows) that appear in each relational table.

## Case Study

Now we will present a preliminary case using selective Cox2 inhibitors, which are a class of pain relief drugs that were found to increase a patient’s risk of having a myocardial infarction (MI) (i.e., a heart attack) (Kearney et al. 2006). We will look at the following two patient populations:

**Population 1 ( $P_1$ ):** The positive examples consist of patients who had a MI. To create a set of negative examples, we selected patients who did not have an MI. None of the patients in this data set were prescribed a selective Cox2 inhibitor.

**Population 2 ( $P_2$ ):** The positive examples consist of patients who had an MI after taking a selective Cox2 inhibitor. To create a set of negative examples, we took patients who were prescribed a selective Cox2 inhibitor and did not have a subsequent MI.

We matched the two populations, population 1 and population 2, on age and gender and were careful to match positive examples with positive examples, and negatives with negatives. The negatives in population 2 were selected to be *all* individuals in the database who had a record of taking Cox2ib, but had no subsequent MI. Thus the prevalence of MI in the Cox2ib population is maintained in the data set, and thus there is a distribution skew between negatives and positives. Due to the large skew in the class distribution, we employ Precision-Recall curves for evaluation, and the areas under them (AUC-PR), rather than ROC curves and AU-CROC. It is possible to build a predictive model for each population, and each can be viewed as a separate but similar learning task. To be able to compare AUC-PR results meaningfully between the two tasks, our matching ensures that both populations have the same skew. One limitation is that, as a result, the prevalence of MI in population 1 is the same as in population 2, roughly double what one would normally see in a general population; nevertheless, there is no way to address this without either having the wrong skew, or propensity, in population 2 or comparing AUC-PRs for different skews. PR curves and the areas under them are known to be sensitive to skew, unlike ROC curves (Boyd et al. 2012).

We designed two simple experiments to try to assess the overlap between these two first tasks. In both cases, we perform stratified ten-fold cross validation to estimate the generalization ability of the models.

**Experiment 1.** In our first experiment, we use SAYU to learn two different models to distinguish between patients that suffered a MI (positive examples) and those that did not (negative examples). The essential difference between the two models is whether the patient took a selective Cox2 inhibitor:

- Model  $M_1$  is trained on patients who did not take a selective Cox2 inhibitor.
- Model  $M_2$  is trained using patients who were prescribed a selective Cox2 inhibitor.

Then we evaluated  $M_1$  and  $M_2$ ’s performance on *both* patient populations. That is,  $M_1$  is used to predict MI for patients who have not been prescribed selective Cox2 inhibitors and those who have. We evaluate  $M_2$  in a similar manner. If one of the models performs well on the other patient population, this would provide evidence that the learned models are picking up general signals for predicting MI rather than MI as an adverse event associated with selective Cox2 inhibitors.

**Experiment 2.** Our second experiment builds on our first. This can be viewed as a theory refinement or transfer learning experiment. At a high-level, the approach works as follows. First, SAYU is employed to learn a set of features using one patient population. Second, this set of features is transferred to the other population where SAYU uses this learned model as a starting point. (Note that SAYU normally starts with an empty feature set.) SAYU uses the data from the second patient population to learn additional features as well as the structure of the statistical model.

Specifically, we learn the following two models:

- Model  $M'_1$  is learned the following manner. It starts by using  $M_2$  (from the first experiment) as the initial model. Then, SAYU refines this model by using the data about patients who did not take a selective Cox2 inhibitor.
- Model  $M'_2$  is learned the following manner. It starts by using  $M_2$  (from the first experiment) as the initial model. Then, SAYU refines this model by using the data about patients who were prescribed a selective Cox2 inhibitor.

Then we perform the same evaluation as before.  $M'_1$  and  $M'_2$ ’s performance is measured on *both* patient populations. Note that we take care to ensure that the test partitions of the populations are never used during training.  $M'_1$  is used to predict MI for patients who have not been prescribed selective Cox2 inhibitors and those that have. We evaluate  $M'_2$  in a similar manner.

## Results for Experiment 1

Table 2 reports results for this experiment. First, consider the cross-validated AUC-PR when the models are trained and tested on the same patient population. Here, we see that the cross-validated AUC-PRs are roughly equivalent. Using a two-tailed unpaired  $t$ -test, we found that there is no significant difference (p-value 0.50).<sup>3</sup> That is, there seems to be

<sup>3</sup>Here the test sets are different, so we use the unpaired test.

Table 1: Characteristics of Patient Populations.

	Cox2ib ( $P_2$ )	No Cox2ib ( $P_1$ )
Positive examples	184	184
Negative examples	1,776	1,776
Unique medications	2,489	2,093
Unique diagnoses	7,221	5,838
Unique observations	1,608	1,492
Medicine table size	1,345,547	286,701
Disease table size	1,136,755	336,708
Observation table size	1,497,693	907,802

Table 2: Results for Experiment One.

Train Population	Testing Population	AUC-PR
No Cox2ib	No Cox2ib	0.560
Cox2ib	Cox2ib	0.602
No Cox2ib	Cox2ib	0.478
Cox2ib	No Cox2ib	0.401

Table 3: Results for Experiment 2.

Train Population	Testing Population	AUC-PR
Cox2ib + No Cox2ib	No Cox2ib	0.601
No Cox2ib + Cox2ib	Cox2ib	0.611
Cox2ib + No Cox2ib	Cox2ib	0.534
No Cox2ib + Cox2ib	No Cox2ib	0.578

no significant difference on our ability to predict MI regardless of whether we condition the patient population based on whether or not a selective Cox2 inhibitor was prescribed. This is interesting as provides some evidence that, on this dataset, the difficulty of these prediction tasks is not substantially different (i.e., one is not a fundamentally more challenging problem than the other).

However, when we use  $M_1$  to predict on patients that were prescribed selective Cox2 inhibitors, the performance drops by 20%. This corresponds to a significant decrease in performance (p-value 0.022) according to a two-tailed paired  $t$ -test.<sup>4</sup> More dramatically, when we use  $M_2$  to predict on the patients that were not prescribed selective Cox2 inhibitors, the performance degrades by 28%. These results can be seen graphically in Figure 1. Again, this difference is significant (p-value 0.0041) according to a two-tailed paired  $t$ -test. This finding is interesting as it gives some evidence that the learned models are tailored to a specific prediction task. Note that the search space of possible models is similar for both problems. The models are picking up on some joint signal, as the models are doing better than random guessing, which would correspond to AUC-PR of 0.094 on these tasks. However, in each task there clearly seems task-dependent structure that, when discovered, leads to significantly better performance.

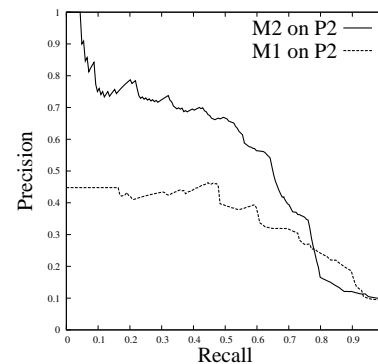
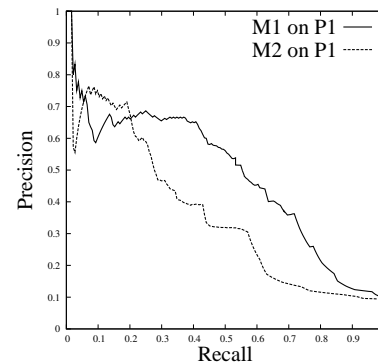


Figure 1: Precision-recall curves for the  $M_1$  and  $M_2$  models on  $P_1$  and  $P_2$ .

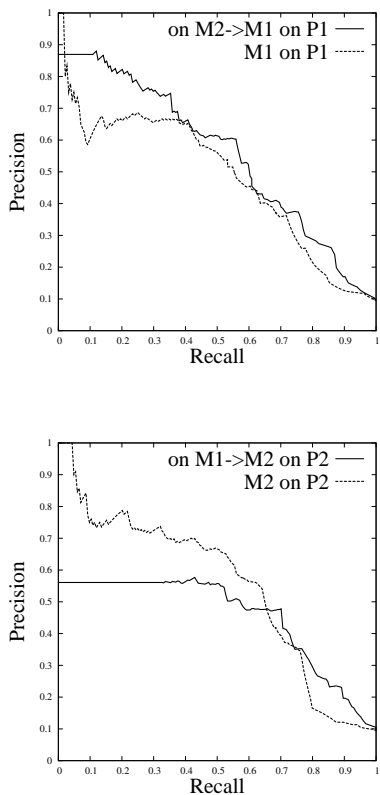


Figure 2: Precision-recall curves for the  $M_1$  and  $M_2$  models on  $P_1$  and  $P_2$ .

## Results for Experiment 2

Table 3 reports results for this experiment. First, we compare  $M'_1$  to  $M_1$  and  $M'_2$  to  $M_2$ . Remember that the models  $M'$  are obtained in two steps: first a model is trained on a different population and second the model is reinforced with new rules, designed to fit the target population. Although the average result in both cases is slightly better than just training on the target population, using a two-tailed paired  $t$ -test, we found there is no significant benefit by training first on the other population. It is not surprising that the results improve, as we are effectively using more data to train the model. It is interesting that the incorporating the extra data has limited benefit. This is again some evidence that the task specific data is crucial for achieving good performance. Figure 2 shows the PR curves for these comparisons.

We then apply this model in the other population: so the final Cox2ib model is evaluated on the non Cox2ib population, and vice-versa. Unsurprisingly, the results improve compared to experiment 1 as the model incorporates structure from both populations, even if it is fine-tuned for the wrong population. It is interesting to note that the improvement in the Cox2ib population (33%) is much greater than for the non Cox2ib population (20%).

<sup>4</sup>Here the test sets are identical, so we use the paired test.

## Conclusions and Future Work

The second PR curve in Figure 1 indicates that if we want a very high recall—we want to catch everyone who will have an MI on a Cox2 inhibitor (Cox2ib)—then we would do better to use the general MI model than the Cox2ib-specific model. Nevertheless, the large gap between the two models at higher precision indicates that the Cox2ib-specific model ranks at highest risk many patients not considered high risk by the general MI model. Potentially we could greatly improve safety and reduce MI, with relatively low impact on overall drug use, if we just denied Cox2ib drugs to the very highest risk patients under the Cox2ib-specific model, for example operating either at precision of roughly 1.0 and recall of roughly 0.05 or operating at precision of roughly 0.7 and recall of roughly 0.4 on the Cox2ib-specific model’s PR curve. Nevertheless, before such a method could be widely implemented, many more issues need to be examined.

One issue is that this study used the entire class of Cox2 inhibitors. How would the curves differ for specific drugs, and in particular how would they differ for drugs still on the market vs. for drugs pulled from the market already? For drugs still on the market, it makes sense to operate in the high precision, low recall range, where we can improve safety without greatly limiting who gets the drugs. If one wanted to take the more dramatic step of using predictive models to bring a drug back to market that had been pulled,<sup>5</sup> it would be necessary to operate in the high recall range to ensure that returning a drug to market did not decrease health and safety.

A second issue is that PR curves alone do not tell the entire evaluation story. It would also be important to see actual numbers of patients estimated to be affected. One useful number from the medical community is “number needed to treat” (NNT). For example, if we implement the predictive model at precision roughly 1.0 and recall roughly 0.05 to deny the drug to a small subset of high-risk individuals, how many patients would need to be seen in order to avoid one MI?

A third issue is that, while in machine learning we are usually happy with a sound cross-validation methodology to compare approaches, in the medical community a prospective clinical trial would be necessary to determine whether use of the predictive model in the clinical setting actually yields health benefits. Hence further study of this modeling task is needed, with the aim of eventually taking the resulting predictive model into the clinical setting in a prospective clinical trial. Some patients with indications for an already-on-the-market Cox2 inhibitor would be prescribed the drug (or not) under the normal procedures, while others would have their data run through the predictive model first before being prescribed the drug—the proposed new procedure. Patients would be followed for several years to determine whether the new procedure reduces MI risk.

In summary this paper presents an initial retrospective study to see how accurately MI can be predicted in a general

<sup>5</sup>This might be desirable for the sake of patients for whom only the pulled drug was efficacious and who are deemed at low risk of the ADR.

population and in a population exposed to Cox2 inhibitors. The results indicate that machine learning can identify with reasonable accuracy Cox2ib-induced risk of myocardial infarction. These promising results encourage further work into models of Cox2ib-induced MI risk, other models of drug-induced risk for negative health outcomes, and other health prediction tasks. From a technical perspective, investigating applying uplift modeling techniques seems like a promising direction to pursue (Jaskowski and Jaroszewicz 2012). This further work needs to include not only improving methods in machine learning and methods in evaluating machine learning results, but also improved methods in analysis of EHR data, causal inference, and clinical trials of predictive models in health care.

### Acknowledgments

JD is partially supported by the research fund KU Leuven (CREA/11/015 and OT/11/051), and EU FP7 Marie Curie Career Integration Grant (#294068). VSC is funded by the ERDF through the Progr. COMPETE and by the Portuguese Gov. through FCT-Fund. for Science and Tech., proj. LEAP ref. PTDC/EIA-CCO/112158/2009 and FCOMP-01-0124-FEDER-015008, and proj. ADE ref. PTDC/EIA-EIA/121686/2010 and FCOMP-01-0124-FEDER-020575. The authors gratefully acknowledge the support of NIGMS grant R01GM097618.

### References

- Boyd, K.; Costa, V. S.; Davis, J.; and Page, D. 2012. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings for the 29th International Conference on Machine Learning*.
- Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. In Cohen, W. W., and Moore, A., eds., *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, 233–240. ACM.
- Davis, J.; Burnside, E. S.; de Castro Dutra, I.; Page, D.; and Santos Costa, V. 2005. An integrated approach to learning bayesian networks of rules. In Gama, J.; Camacho, R.; Brazdil, P.; Jorge, A.; and Torgo, L., eds., *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, 84–95. Springer.
- Davis, J.; Ong, I.; Struyf, J.; Burnside, E.; Page, D.; and Costa, V. S. 2007. Change of representation for statistical relational learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2719–2726.
- Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian networks classifiers. *Machine Learning* 29:131–163.
- Getoor, L., and Taskar, B., eds. 2007. *An Introduction to Statistical Relational Learning*. MIT Press.
- Jaskowski, M., and Jaroszewicz, S. 2012. Uplift modeling for clinical trial data. In *Proceedings of ICML 2012 Workshop on Machine Learning for Clinical Data Analysis*.
- Kearney, P.; Baigent, C.; Godwin, J.; Halls, H.; Emberson, J.; and Patrono, C. 2006. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? meta-analysis of randomised trials. *BMJ* 332:1302–1308.
- Muggleton, S. 1995. Inverse entailment and Progol. *New Generation Computing* 13:245–286.