Carnegie Mellon University
Software Engineering Institute

MACHINE LEARNING IN CYBERSECURITY: A GUIDE

Jonathan M Spring, Joshua Fallon, Leigh Metcalf

Feb 2019

Executive Summary

Decision-makers should ask certain questions before employing machine-learning (ML) or artificial intelligence (AI) solutions – and receive satisfactory answers. This document suggests important questions when employing ML or AI in cybersecurity and outlines what a satisfactory answer should contain. We focus on questions about quality and usefulness. The questions we discuss are:

1. What are you trying to find out?
2. What information is needed to answer the target question?
3. How do you anticipate that the ML/AI tool will address that question?
4. Is the design of the ML/AI tool robust to the well-known attacks against ML/AI in our adversarial, cybersecurity environment?
5. How can the input data's bias be managed?
6. Does the evaluation of the ML/AI tool properly account for well-known study design errors and biases?
7. What alternative tools have you considered? What are the advantages and disadvantages of each?

Introduction

Machine learning (ML) is a set of statistical tools to infer models of data. Artificial intelligence (AI) does not mean the sci-fi dream of a thinking robot, but rather coupling a classification tool such as ML with a controller that can act based on the classification. ML and AI are becoming popular tools for problems within cybersecurity. This paper will not offer a tutorial on either cybersecurity or ML.[1] Our goal is to provide managers and decision-makers with practical questions about the quality and usefulness of an ML/AI tool, and the shape of a satisfactory answer to each. We will not assume any

[1] For an intro to ML, the Wikipedia page is reasonably accessible (https://en.wikipedia.org/wiki/Machine_learning). For fuller details, see Andrew Moore's tutorial page at https://www.cs.cmu.edu/~awm/tutorials.html.

knowledge about ML/AI, but we do assume an understanding of cybersecurity concepts like confidentiality, integrity, and availability as well as related concepts like what malicious software (malware) is.[2]

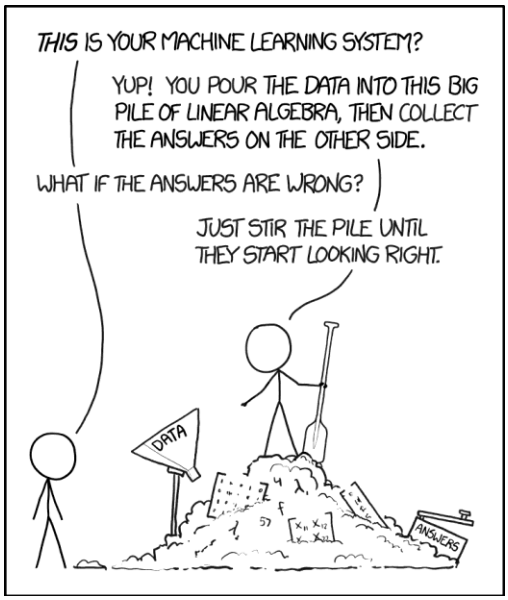

Figure 1: "Machine Learning" by Randall Munroe (CC BY-NC 2.5)

ML/AI is a tool, or more accurately a suite of tools.[3] The appropriate question is therefore whether a tool is fit for its intended purpose. The questions we highlight in the following sections expand this general question out to some more tractable questions. Some of these questions may not be specific to cybersecurity applications of ML/AI tools, but a cybersecurity application does affect the properties of a satisfactory answer for all the questions. One other important idea is that we are not trying to find or define optimal or best tools within cybersecurity, but rather satisfactory or good-enough tools.

A note on good evidence. The answers to all these questions require evidence. Therefore, it is worth commenting on some general features of good evidence. Evidence should be collected via structured observations that are designed to reduce mistakes and biases; the evidence source needs to contain information relevant to the question. The methods of the sciences are a good source for advice.[4] In cybersecurity, evidence should be interpreted knowing an adversary may be influencing the decision-making process.[5]

What are you trying to find out?

The first question to ask about an ML/AI tool is "What are you trying to find out?" This question is important for at least two reasons. First, without knowing the purpose of the tool, we cannot evaluate the other questions. ML/AI tools do not generate questions; a useful question must be posed to the tool. Knowing the question is not the same as proposing metrics and measurements.

An example question is: Is a given attack on the organization similar or related to certain prior attacks? One common unhelpful example might be "is this computer behavior weird?" Anomalous behavior,

---

[2] A two-page summary is available from: US-CERT, Introduction to Information Security, 2008 (https://www.us-cert.gov/sites/default/files/publications/infosecuritybasics.pdf). For a more thorough intro, see for example Anderson RJ. Security engineering: a guide to building dependable distributed systems. John Wiley & Sons; 2008.

[3] The dictionary definition of a tool is something "with which some operation is performed; a means of effecting something; an instrument" ("tool, n." *OED Online*, Oxford University Press, December 2018, www.oed.com/view/Entry/203258. Accessed 17 December 2018).

[4] Spring JM, Moore T, and Pym D. Practicing a Science of Security. In: Proc. 2017 New Security Paradigms Workshop, Santa Cruz, CA, USA, October 1–4, 2017.

[5] Horneman A. How to think like an analyst. July 17, 2017. https://insights.sei.cmu.edu/sei_blog/2017/07/how-to-think-like-an-analyst.html

whether it is emails, network traffic, or software operations, need not imply a security violation. If a tool is going to be a cybersecurity tool, it should ask a cybersecurity question, not a question on common sociological uses of computers.

Do you have the necessary information?

ML/AI tools can find connections within data – sometimes, surprising connections. But they cannot create something from nothing. This question is not meant to require an enumeration of just the right data fields. It is rather intended to make sure the right kind of information is available. For example, the security of an information system is relative to the security policy of the organization. If an ML/AI tool is going to make a detection decision as to whether certain code is malicious, it should have access to data about the security policy.[6] Computer code is not a security policy, and software may behave in ways that violate one organization's policy but not another's. Or certain software at one time may violate an organization's policy but at other times and contexts may not. If your question of interest was correlating code snippets with a specific organization's well-defined security policy, within a margin of error, then this is probably a reasonable ML/AI task. An example question of interest that would fail this test is "is a given piece of software malicious in general?" because the input data does not have the right type of information (namely, about the human concept of maliciousness).

How can the result be explained and understood?

The third question is to ask "How do you anticipate that the ML/AI tool will address that question?" The type of question influences what sort of tools yield appropriate answers. The question words (who, what, where, why, when, how) help indicate what type of tool can answer a question. Questions about what exists or what is, or how observing something changes beliefs about something else, are associative questions. The formalism within ML/AI tools is well-suited to answer these sorts of questions. If the question at hand is something like "is a given email spam?" then all is well: move on to the next section. If the question you want to answer is a question about intervening on your system (what if…), or a question about why something happened the way that it did, then ML/AI are not well-suited to your question.[7] Better tools might be deployable logical reasoning,[8] or structured general knowledge[9] to support counterfactual reasoning, as cybersecurity often needs to answer "what if…" and why questions. As one example, remediation during incident response should include an explanation of why any given

---

[6] For definitions of terms such as "security policy" see: Shirey R. Internet security glossary, version 2. RFC 4949. 2007.

[7] Pearl J. The Seven Tools of Causal Inference with Reflections on Machine Learning. http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf

[8] See for example: Klein G, Elphinstone K, Heiser G, Andronick J, Cock D, Derrin P, Elkaduwe D, Engelhardt K, Kolanski R, Norrish M, Sewell T. seL4: Formal verification of an OS kernel. Symposium on Operating Systems Principles (SIGOPS). 2009 Oct (pp. 207-220). ACM.

[9] Spring JM, Illari P. Building general knowledge of mechanisms in information security. Philosophy & Technology. 2018.

change to the system will help remove the adversary's presence on the system. It is vital to have a strategy to bridge the gap between what questions cybersecurity often asks and what questions ML/AI tools are well-suited to answer.

The ML/AI tool will serve as or assist a human expert in answering the question at hand. To understand whether the tool will be fit for purpose, it is worth considering how we evaluate expertise generally. A layperson questioning an expert will receive advice or an answer to their question from the expert, or perhaps the expert will perform a service in response to the layperson's request. Our challenge is to determine how the layperson may expect to have the result explained and understand the answer without becoming an expert themselves. Furthermore, different stakeholder communities have different expectations of explanations related to ML/AI tools.[10]

With human experts, we often expect one of two things: an immediate demonstration of success or for them to explain their decision-making process. The former option requires that the service or expertise be immediately testable. For example, if a technician fixes a broken device (a car or clock, for example) then the proof of their expertise is immediate – the device either works or not. But if the layperson wants to know something which is not immediate testable, for example whether the car will continue to work based on this fix for the next five years, more is needed. In general, the social system that seems to work is that the expert should provide an explanation. This explanation need not necessarily be to the layperson, if it occurs transparently among other experts acting in good faith.[11]

Therefore, an adequate answer to the question should indicate whether the results of the ML/AI tool are immediately testable or not. If the results are immediately testable, then the result contributes to understanding of the cybersecurity situation via these tests. Whether the tests are robust falls under the following questions. However, if the ML/AI tool is going make decisions for which the impact will not be felt until relatively far in the future, the decision-making process will need to be evaluated as an expert explanation. Within cybersecurity, whether an email is spam is immediately testable, for example, because users can quickly provide feedback. Whether a particular piece of software or element of network traffic is indicative of an intrusion into a computer is probably not immediately correctable or open to feedback – most intrusions are not detected for months after they start.[12]

Explanations of ML/AI tools are the subject of a research area known as explainable AI. The important aspect of a good answer from an expert about their decision-making is whether it is relevant to helping the layperson understand why the expert's choice is reliable. Explainable AI tends to focus on details of the ML apparatus. Returning to the example of the car technician, consider asking for an explanation

---

[10] Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S. Stakeholders in Explainable AI. arXiv preprint arXiv:1810.00184. 2018 Sep 29.

[11] For a discussion of the topic of expertise and further references, see: Douglas H. How can the Public Assess Expertise? JBS Haldane Lecture, UCL, Jan 2018. https://www.youtube.com/watch?v=cuB06iZ8-sM

[12] See the Verizon Data Breach Investigation Report series, which documents that time from initial intrusion to detection is most often measured in months (see 2009 to 2018 reports). https://enterprise.verizon.com/resources/reports/dbir/

for why the repair of the transmission is good enough. If a neuroscientist tells you that electroencephalogram readings were taken of the technician's brain during the repair and that these readings indicate appropriate hippocampal functioning related to long-term memory, this explanation is not strictly false. It is also, in some sense, encouraging. But it is not really the right kind of explanation to help you answer your question. It is not an explanation at the correct level.[13]

A good explanation might take the phenomenon of interest, as identified by the first question, and break it up into entities and activities and show how their interaction and organization is responsible for the phenomenon. This kind of explanation demonstrates how a change in one part might impact the problem at hand.[14] For example, research into spam-advertising revenue streams suggested the weak-point in the interaction was the Visa payment processor.[15] Even such a hypothetical impact provides some starting point for forming hypotheses and gathering evidence. If the results from an ML/AI tool cannot be explained in a similar way, then the tool cannot be reliably evaluated with respect to the relevant questions.

How is the ML/AI tool robust against manipulation and attack?

ML/AI tools have well-known vulnerabilities to adversary manipulation. So a natural question in a cybersecurity context, which explicitly includes adversaries, is "how are the design and deployment of the ML/AI tool robust to the well-known attacks?" This is not an abstract question about all possible attacks. There are well-documented classes of attacks; attacks during training-time and classification-time, attacks on integrity, confidentiality, and privacy.[16] Cybersecurity cannot and is not about preventing all attacks, but any tool deployed in a cybersecurity context should have an explicit risk assessment. For ML/AI tools, this means an assessment of vulnerability to each of these various classes of attack. A good answer to this question should include both robust design of the ML/AI algorithm and tool itself, as well as how attacks that bypass or manipulate the deployed ML/AI tool can be detected and mitigated.

As a poignant example of attacks against ML/AI tools, consider the case of self-driving cars. Such cars use ML tools to identify street signs (among other things). By intentionally manipulating a small section of a stop sign with a purpose-designed sticker, an adversary can make these operational ML/AI tools reliably misclassify a stop sign as a 45-mile-per-hour speed limit sign.[17] Cybersecurity tools are constantly exposed to adversary input, so any cybersecurity ML/AI tools need to take this threat seriously.

---

[13] By "level" we mean mechanistic level of explanation as described in: Craver C. Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford University Press; 2007.

[14] There is a large literature on these topics in the physical, social, and life sciences, see Glennan S, Illari PM, editors. The Routledge handbook of mechanisms and mechanical philosophy. Taylor & Francis; 2017.

[15] Kanich C, Weaver N, McCoy D, Halvorson T, Kreibich C, Levchenko K, Paxson V, Voelker GM, Savage S. Show Me the Money: Characterizing Spam-advertised Revenue. USENIX Security Symposium 2011 Aug 8.

[16] Papernot N, McDaniel P, Sinha A, Wellman MP. SoK: Security and privacy in machine learning. IEEE European Symposium on Security and Privacy (EuroS&P) 2018 Apr 24 (pp. 399-414). IEEE.

[17] Evtimov I, Eykholt K, Fernandes E, Kohno T, Li B, Prakash A, Rahmati A, Song D. Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945. 2017 Jul 27.

How you adequately guarded against bias in input data?

Input data can be biased in several ways. Some problems are not specific to security, but cybersecurity does present some unique considerations about input data that may require special attention. The first element is latent bias in the data. This problem has been vividly demonstrated in ML/AI tools used to advise on criminal sentencing in the US. African Americans are four times more likely to be arrested for drug charges than are white Americans, despite similar rates of usage. ML/AI tools that have been trained to provide information on recidivism have re-learned this bias in the criminal justice system even when race is specifically excluded as a data input.[18] Underlying human bias is frequently re-learned by such systems.[19]

In a cybersecurity context, mitigating input data bias helps ensure both the general quality of the ML/AI tool as well as the fair impact on the users of the system. A good answer to the question of input data bias will include assurances about the following five aspects of data quality:[20]

1. Representation – all relevant subjects are proportionally represented in the data. For example, the ratio of benign to malicious elements is realistic, and cultural bias in what sorts of items count as "benign" is mitigated.
2. Protection – confounding factors are not mis-learned as proxies for sensitive characteristics.
3. Stewardship –the relevant communities impacted by and producing the data are engaged.
4. Authenticity – the features of the training data are faithful to the application environment.
5. Resiliency – adversary access to the training or benchmarking data will not allow the adversary to trivially interfere with the ML/AI tool; or the tool does not rely on static input data.

Is the ML/AI tool evaluated adequately?

The next question to ask is "Does the evaluation of the ML/AI tool properly account for well-known study design errors and biases?" At this point, you have established the question of interest, the intelligibility of potential answers, the robustness of the method to attacks, and the appropriateness of the input data. So the remaining question is about whether the tool, once built, meets your needs.

---

[18] Angwin, J, Larson J, Mattu S, Kirchner L. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. May 23, 2016. ProPublica.

[19] Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017 Apr 14;356(6334):183-186. https://doi.org/10.1126/science.aal4230

[20] Numbers one through four are adapted from: Polonski V. AI is convicting criminals and determining jail time, but is it fair? World Economic Forum. Nov 19, 2018. For other ethical considerations on collecting data in cybersecurity studies, see: Dittrich D, Kenneally E. The Menlo Report: Ethical principles guiding information and communication technology research. US Department of Homeland Security. 2012 Aug.

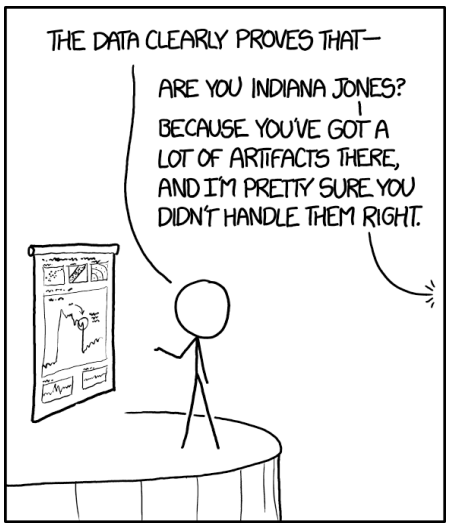

Figure 2: "Artifacts" by Randall Munroe (CC BY-NC 2.5)

There are various common errors to evaluation of tools, both in and out of cybersecurity, with and without ML/AI being involved. We will discuss three common hazards to evaluation of ML/AI tools in cybersecurity. In summary, make sure any evaluation (1) accounts for the base-rate at which the evaluated property occurs; (2) fully understands the population being evaluated, and; (3) properly accounts for missing evidence.

The impact of errors on operators who manage alerts generally, such as nuclear power plant operators or computer security incident responders, is due to *alarm error*. This is a measure of the probability that an alert received by an operator is a false alarm. Alarm error differs from the false alarm rate, often called the false positive rate, usually reported about ML/AI tools, because alarm error takes into account the base rate of occurrence as well as the false positive rate. Many relevant cybersecurity events are relatively rare. For example, most software instances are not malicious. The effect of a low base-rate of occurrence, such as this, is that seemingly small false positive rates result in alarm error rates that overwhelm the operator. Consider an example with a 1% test error (1% false positive and false negative rates), and a base-rate for the item of interest of one in 10,000, with the test applied to a population of 10,000. Despite a 1% error rate in the test, about 99% of the alerts from the test will be alarm errors.[21] The fact that there are so many more benign cases in the population drastically impacts the results.

It is common for evaluation of a tool to be done on some subset of the relevant population. An example of bias in misunderstanding how this population represents the population as a whole is called *survivorship bias* because it was first studied in evaluating new armor for returning military aircraft.[22] The problem was to estimate what vulnerable parts of the plane most often lead to a plane being shot down, with only planes that were not shot down available for observation. The general problem also relates to cybersecurity – commonly, we want to estimate something unobservable, such as how many intrusion events an organization did not detect. There are statistical methods for such evaluations. However, they must be selected and applied carefully and with good reasoning. The main pitfall that needs to be avoided is treating an observed or evaluated population as representative when it is systematically not representative. With the airplane example, the right answer was to put more armor where the returning planes were not shot. The planes that did not return more likely took fatal damage in the places the successful planes remained unscathed. Similarly, taking as input data all the intrusions an organization

---

[21] Axelsson S. The base-rate fallacy and the difficulty of intrusion detection. ACM Transactions on Information and System Security (TISSEC). 2000 Aug 1;3(3):186-205.

[22] Mangel M, Samaniego FJ. Abraham Wald's Work on Aircraft Survivability. J American Statistical Association, 1984 79:386, 259-267, DOI: 10.1080/01621459.1984.10478038

knows about is an unlikely strategy for evaluating a tool's ability to detect intrusions the organization does not know about.

The final common error to guard against during evaluation of ML/AI tools is misunderstanding absent evidence. Intelligence analysis has a long history of addressing this challenge,[23] on which cybersecurity should draw. It is worth quoting Heuer's advice on overcoming this problem directly as advice on what any answer to the question of evaluation quality should include: "[I]dentify explicitly those relevant variables on which information is lacking, consider alternative hypotheses concerning the status of these variables, and then modify … judgement accordingly. [Also] consider whether the absence of information is normal or is itself an indicator of unusual activity or inactivity."

These are not the only considerations about evaluation of tools. However, the above considerations form a solid basis for an answer to this question.

What are the advantages and disadvantages of the tool?

The final question is about the relationship between the ML/AI tool you are considering and other tools. Answers to the previous questions for multiple tools, whether ML/AI-based or not, give a grounding to compare options. There are also business considerations that might make one tool more suitable over another. Estimating costs of development, deployment, and maintenance are important problems of their own,[24] which we will not focus on here. However, be sure to consider advantages and disadvantages at various stages in the tool's lifecycle. In the case of cybersecurity, it is particularly relevant to understand the threat lifecycle and how a tool can be updated when an adversary learns how to subvert it. It is undesirable if one week of adversary effort takes a 3-month tool redevelopment to counter, for example. More generally, try to predict the adversary's response to the tool and whether that response puts them in a better or worse position.[25]

Conclusion

Machine learning (ML) and artificial intelligence (AI), like any tools, should be designed such that they are fit for their intended purpose. We have provided five questions a manager or decision maker might ask of any ML/AI tool to be employed in cybersecurity, and suggested some desirable features of answers. These questions, and the features of the answers, can be summarized as follows.

1. What are you trying to find out?

---

[23] Heuer RJ. Psychology of intelligence analysis. US CIA; 1999. p 119, http://www.dtic.mil/get-tr-doc/pdf?AD=ADA500078

[24] See for example the publications at https://insights.sei.cmu.edu/sei_blog/software-cost-estimates/

[25] Spring JM, Stoner E. CND Equities Strategy. CERTCC-2015-40. Pittsburgh, PA. Jul 2015.

- A response should identify a question about cybersecurity topics, such as effects of a specified security policy. In order to successfully apply an ML/AI tool, the question should be about what exists or how observations change beliefs about what exists.

2. What information is needed to answer the target question?

    - A response should demonstrate that the input data encode the same type of information as the answer sought. For example, security policy evaluations such as "maliciousness" are not the same type of data as software instructions.

3. How do you anticipate that the ML/AI tool will address that question?

    - A response should be held to the same standards a human expert would be expected to meet when explaining their decision to a layperson. The response should explain the situation and what we as humans learn from the tool, not how the tool functions.

4. How is the design of the ML/AI tool robust to the well-known attacks against ML/AI?

    - A response should demonstrate protection of the ML/AI tool itself as well as resiliency measures for the environment in which the tool will be deployed. Three important examples of such protection are: (1) robustness in both confidentiality and integrity of the tool; (2) the tool should be resilient against attacks during training and classification; and, (3) evidence that the input data is reliable and representative.

5. How can the input data's bias be managed?

    - A response should consider the five principles of representation, protection, stewardship, authenticity, and resiliency.

6. Does the evaluation of the tool properly account for well-known study design errors and biases?

    - A response should transparently and as fully as possible plan the steps for evaluating a tool. Important considerations include data sources, design of the study, appropriate measures of success such as alarm error, understanding the target population, counterfactual analysis to explore missing evidence, and the extent to which evidence from the evaluation is generalizable.

7. What alternatives have you considered? What are the advantages and disadvantages of each?

    - A fair answer to this question should compare 3 or 4 options using the above six questions. At least one of the alternatives should not be an ML/AI tool. Lifecycle costs, development costs, maintenance and operation costs should all be considered. Since these are cybersecurity tools, also consider the adversary's natural response.

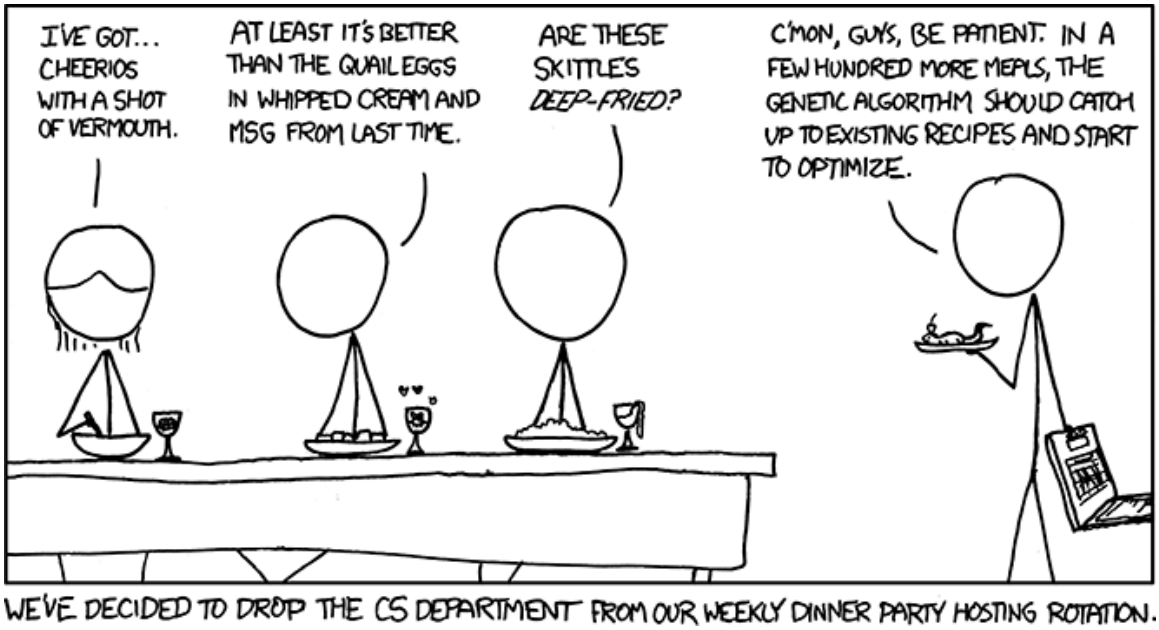

Figure 3: "Recipes" by Randall Munroe (CC BY-NC 2.5)

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu
Email: info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.