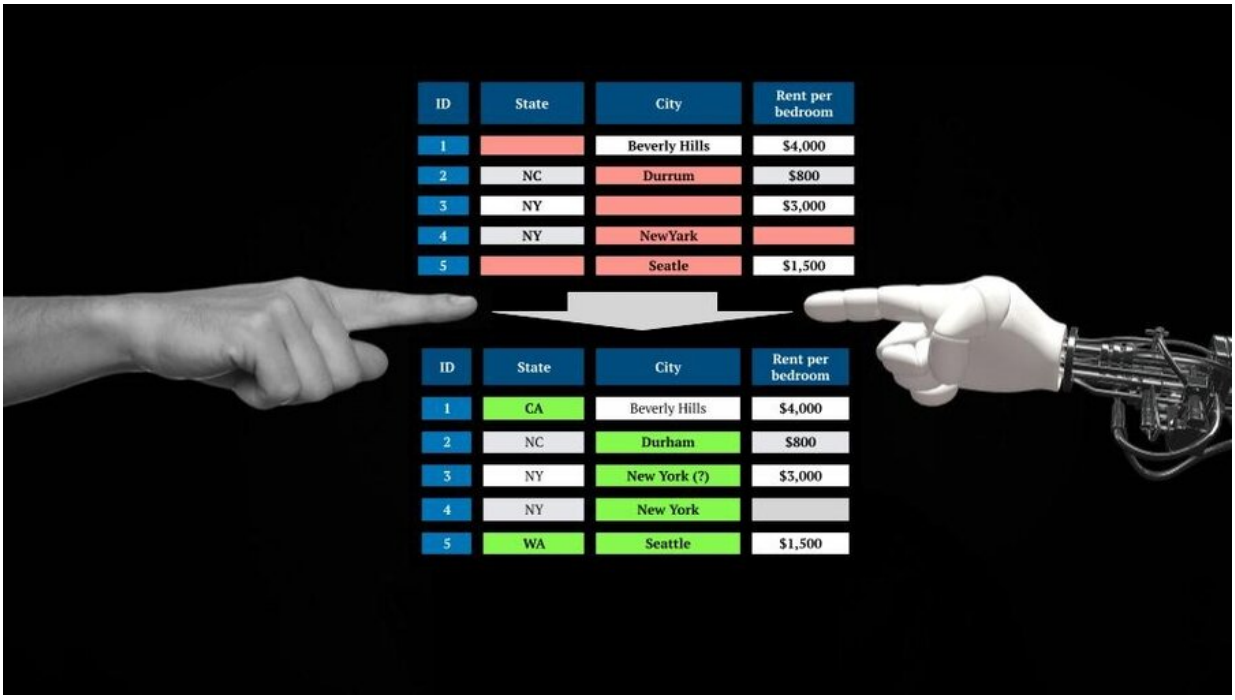
New system cleans messy data tables automatically

May 12 2021



MIT researchers have created a new system that automatically cleans “dirty data” — the typos, duplicates, missing values, misspellings, and inconsistencies dreaded by data analysts, data engineers, and data scientists. Credit: Massachusetts Institute of Technology

MIT researchers have created a new system that automatically cleans "dirty data"— the typos, duplicates, missing values, misspellings, and inconsistencies dreaded by data analysts, data engineers, and data

scientists. The system, called PClean, is the latest in a series of domain-specific probabilistic programming languages written by researchers at the Probabilistic Computing Project that aim to simplify and automate the development of AI applications (others include one for 3D perception via inverse graphics and another for modeling time series and databases).

According to surveys conducted by Anaconda and Figure Eight, data cleaning can take a quarter of a data scientist's time. Automating the task is challenging because different datasets require different types of cleaning, and common-sense judgment calls about objects in the world are often needed (e.g., which of several cities called "Beverly Hills" someone lives in). PClean provides generic common-sense models for these kinds of judgment calls that can be customized to specific databases and types of errors.

PClean uses a knowledge-based approach to automate the data cleaning process: Users encode background knowledge about the database and what sorts of issues might appear. Take, for instance, the problem of cleaning state names in a database of apartment listings. What if someone said they lived in Beverly Hills but left the state column empty? Though there is a well-known Beverly Hills in California, there's also one in Florida, Missouri, and Texas … and there's a neighborhood of Baltimore known as Beverly Hills. How can you know in which the person lives? This is where PClean's expressive scripting language comes in. Users can give PClean background knowledge about the domain and about how data might be corrupted. PClean combines this knowledge via common-sense probabilistic reasoning to come up with the answer. For example, given additional knowledge about typical rents, PClean infers the correct Beverly Hills is in California because of the high cost of rent where the respondent lives.

Alex Lew, the lead author of the paper and a Ph.D. student in the

Department of Electrical Engineering and Computer Science (EECS), says he's most excited that PClean gives a way to enlist help from computers in the same way that people seek help from one another. "When I ask a friend for help with something, it's often easier than asking a computer. That's because in today's dominant programming languages, I have to give step-by-step instructions, which can't assume that the computer has any context about the world or task—or even just common-sense reasoning abilities. With a human, I get to assume all those things," he says. "PClean is a step toward closing that gap. It lets me tell the computer what I know about a problem, encoding the same kind of background knowledge I'd explain to a person helping me clean my data. I can also give PClean hints, tips, and tricks I've already discovered for solving the task faster."

Co-authors are Monica Agrawal, a Ph.D. student in EECS; David Sontag, an associate professor in EECS; and Vikash K. Mansinghka, a principal research scientist in the Department of Brain and Cognitive Sciences.

What innovations allow this to work?

The idea that probabilistic cleaning based on declarative, generative knowledge could potentially deliver much greater accuracy than machine learning was previously suggested in a 2003 paper by Hanna Pasula and others from Stuart Russell's lab at the University of California at Berkeley. "Ensuring data quality is a huge problem in the real world, and almost all existing solutions are ad-hoc, expensive, and error-prone," says Russell, professor of computer science at UC Berkeley. "PClean is the first scalable, well-engineered, general-purpose solution based on generative data modeling, which has to be the right way to go. The results speak for themselves." Co-author Agrawal adds that "existing data cleaning methods are more constrained in their expressiveness, which can be more user-friendly, but at the expense of being quite

limiting. Further, we found that PClean can scale to very large datasets that have unrealistic runtimes under existing systems."

PClean builds on recent progress in probabilistic programming, including a new AI programming model built at MIT's Probabilistic Computing Project that makes it much easier to apply realistic models of human knowledge to interpret data. PClean's repairs are based on Bayesian reasoning, an approach that weighs alternative explanations of ambiguous data by applying probabilities based on prior knowledge to the data at hand. "The ability to make these kinds of uncertain decisions, where we want to tell the computer what kind of things it is likely to see, and have the computer automatically use that in order to figure out what is probably the right answer, is central to probabilistic programming," says Lew.

PClean is the first Bayesian data-cleaning system that can combine domain expertise with common-sense reasoning to automatically clean databases of millions of records. PClean achieves this scale via three innovations. First, PClean's scripting language lets users encode what they know. This yields accurate models, even for complex databases. Second, PClean's inference algorithm uses a two-phase approach, based on processing records one-at-a-time to make informed guesses about how to clean them, then revisiting its judgment calls to fix mistakes. This yields robust, accurate inference results. Third, PClean provides a custom compiler that generates fast inference code. This allows PClean to run on million-record databases with greater speed than multiple competing approaches. "PClean users can give PClean hints about how to reason more effectively about their database, and tune its performance—unlike previous probabilistic programming approaches to data cleaning, which relied primarily on generic inference algorithms that were often too slow or inaccurate," says Mansinghka.

As with all probabilistic programs, the lines of code needed for the tool

to work are many fewer than alternative state-of-the-art options: PClean programs need only about 50 lines of code to outperform benchmarks in terms of accuracy and runtime. For comparison, a simple snake cellphone game takes twice as many lines of code to run, and Minecraft comes in at well over 1 million lines of code.

In their paper, just presented at the 2021 Society for Artificial Intelligence and Statistics conference, the authors show PClean's ability to scale to datasets containing millions of records by using PClean to detect errors and impute missing values in the 2.2 million-row Medicare Physician Compare National dataset. Running for just seven-and-a-half hours, PClean found more than 8,000 errors. The authors then verified by hand (via searches on hospital websites and doctor LinkedIn pages) that for more than 96 percent of them, PClean's proposed fix was correct.

Since PClean is based on Bayesian probability, it can also give calibrated estimates of its uncertainty. "It can maintain multiple hypotheses—give you graded judgments, not just yes/no answers. This builds trust and helps users override PClean when necessary. For example, you can look at a judgment where PClean was uncertain, and tell it the right answer. It can then update the rest of its judgments in light of your feedback," says Mansinghka. "We think there's a lot of potential value in that kind of interactive process that interleaves human judgment with machine judgment. We see PClean as an early example of a new kind of AI system that can be told more of what people know, report when it is uncertain, and reason and interact with people in more useful, human-like ways."

David Pfau, a senior research scientist at DeepMind, noted in a tweet that PClean meets a business need: "When you consider that the vast majority of business data out there is not images of dogs, but entries in relational databases and spreadsheets, it's a wonder that things like this

don't yet have the success that deep learning has."

Benefits, risks, and regulation

PClean makes it cheaper and easier to join messy, inconsistent databases into clean records, without the massive investments in human and software systems that data-centric companies currently rely on. This has potential social benefits—but also risks, among them that PClean may make it cheaper and easier to invade peoples' privacy, and potentially even to de-anonymize them, by joining incomplete information from multiple public sources.

"We ultimately need much stronger data, AI, and privacy regulation, to mitigate these kinds of harms," says Mansinghka. Lew adds, "As compared to machine-learning approaches to data cleaning, PClean might allow for finer-grained regulatory control. For example, PClean can tell us not only that it merged two records as referring to the same person, but also why it did so—and I can come to my own judgment about whether I agree. I can even tell PClean only to consider certain reasons for merging two entries." Unfortunately, the reseachers say, privacy concerns persist no matter how fairly a dataset is cleaned.

Mansinghka and Lew are excited to help people pursue socially beneficial applications. They have been approached by people who want to use PClean to improve the quality of data for journalism and humanitarian applications, such as anticorruption monitoring and consolidating donor records submitted to state boards of elections. Agrawal says she hopes PClean will free up data scientists' time, "to focus on the problems they care about instead of data cleaning. Early feedback and enthusiasm around PClean suggest that this might be the case, which we're excited to hear."

More information: PClean: Bayesian Data Cleaning at Scale with

Domain-Specific Probabilistic Programming.
proceedings.mlr.press/v130/lew21a/lew21a.pdf

Github: github.com/probcomp/PClean

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology