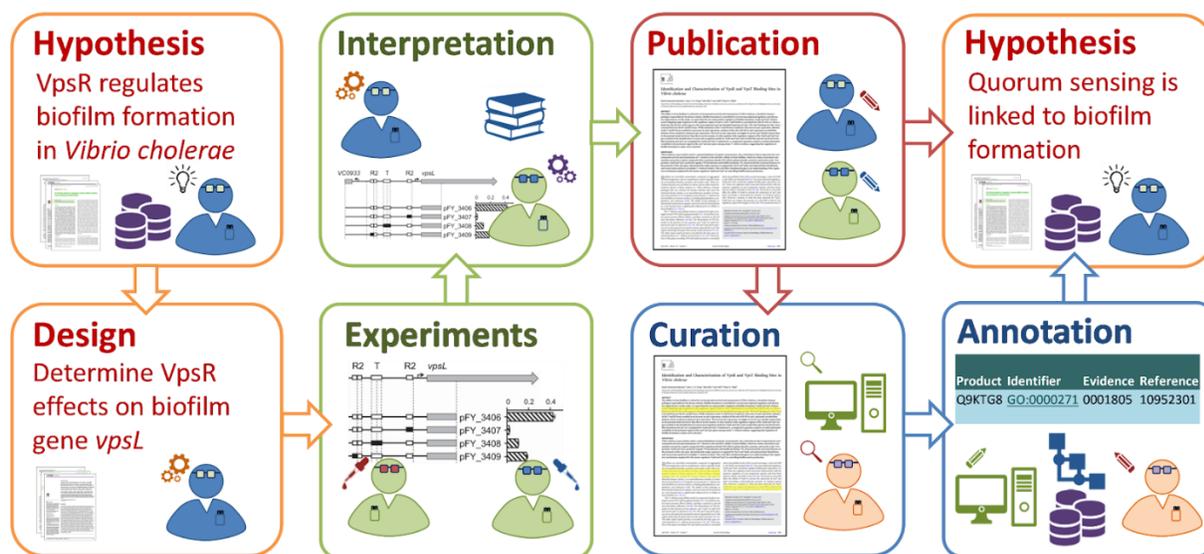


Why annotate?

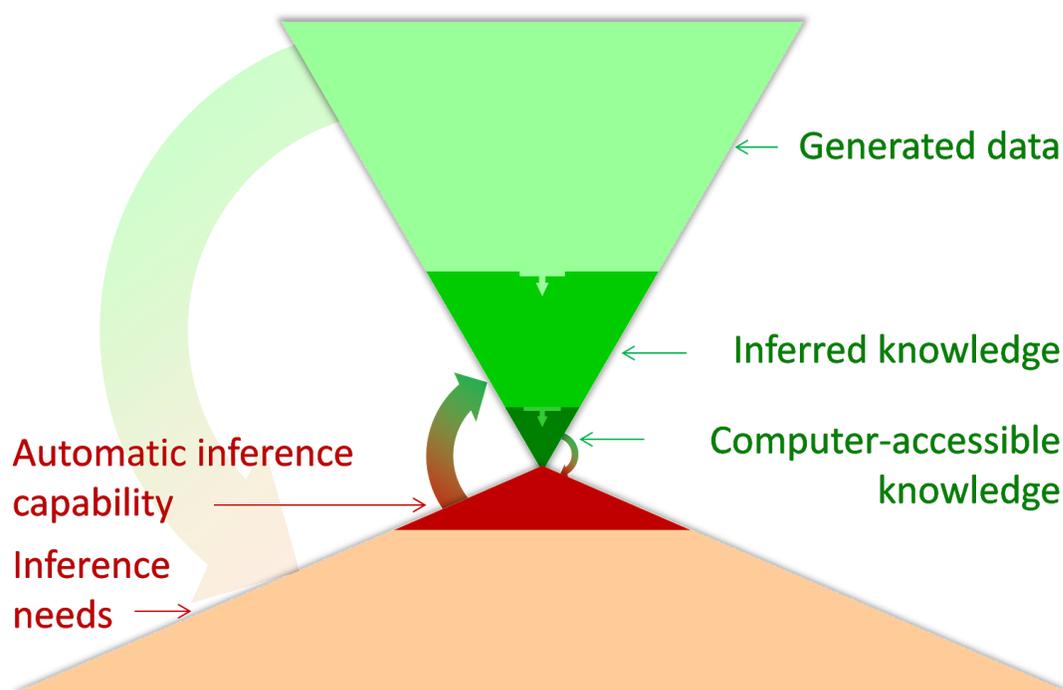
The emergence of high-throughput technologies in the last decade has propelled biology to the forefront of “big data” sciences. Advances in sequencing, sensing, bioinformatics and robotic technologies allow biomedical science practitioners to simultaneously screen the response of cancer cell lines to hundreds of drugs or catalogue the genetic diversity of microorganisms in ocean waters. Open standards and repositories have been created to facilitate the sharing of the unprecedented deluge of data generated by high-throughput biological experiments, and several major initiatives, like the European ELIXIR project or the NIH Big Data to Knowledge (BD2K) platform, are underway to enhance the interoperability, accessibility and reusability of data, as well as the computational methods and experimental protocols associated with its generation. However, the way in which scientific knowledge is disseminated has changed little in the last hundred years (Figure 1). Much like a century ago, science continues to be reported in articles, where information is embedded in a mixture of prose, tables and figures that cannot be readily interpreted by computers. As the Future of Research Communication and e-Scholarship (FoRCe11) manifesto adeptly puts it: “producers and consumers [of science] remain wedded to formats developed in the era of print publication”.



The conventional scientific process workflow, involving publishing and posterior annotation of knowledge in open repositories.

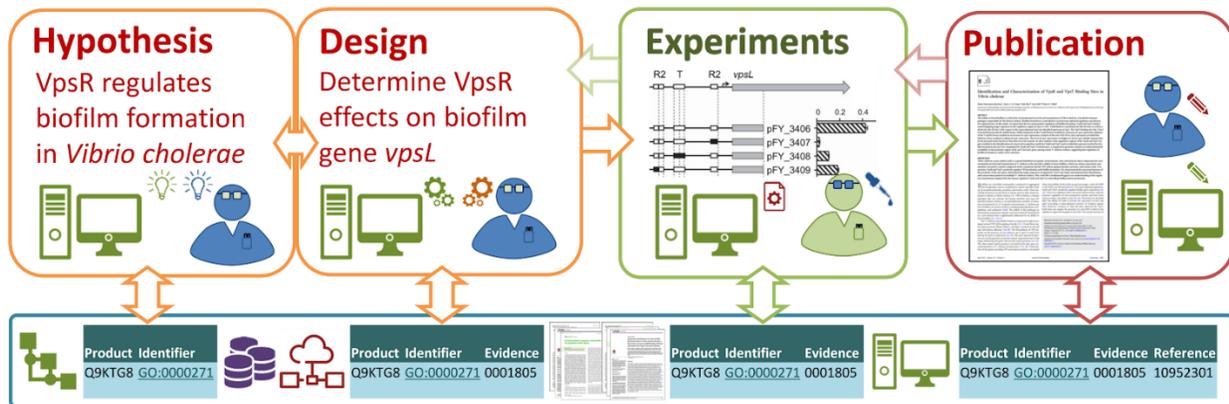
The current reporting model poses a fundamental problem for the progress of science. Scientists have become increasingly dependent on the availability of standardized scientific knowledge in order to infer new knowledge from high-throughput experiments, yet they keep disseminating their results in the form of non-standardized scientific articles. It's kind of silly, really. As a result, vast amounts of time and money must be spent in the manual parsing of scientific articles by expert biocurators and in the development of less accurate methods to extract knowledge from scientific publications and store it in publicly available databases. In

spite of these efforts, only a nominal fraction of the scientific knowledge generated yearly is eventually standardized and deposited in publicly available databases. The lack of standardized knowledge severely limits the ability of automatic inference systems to extract knowledge from high-throughput data, casting serious doubts on the sustainability of the modern scientific enterprise.



The asymmetrical hourglass model of the requirements and availability of computer-accessible knowledge.

Our goal is to develop a system that will help change the way science is communicated by bypassing the costly process of curation. As [Bourne et al.](#) put it: *"In the longer term, we need models that are better aligned with the research life cycle. There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a biocurator extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access."* In essence, we need to move to a model in which the author does the annotation.

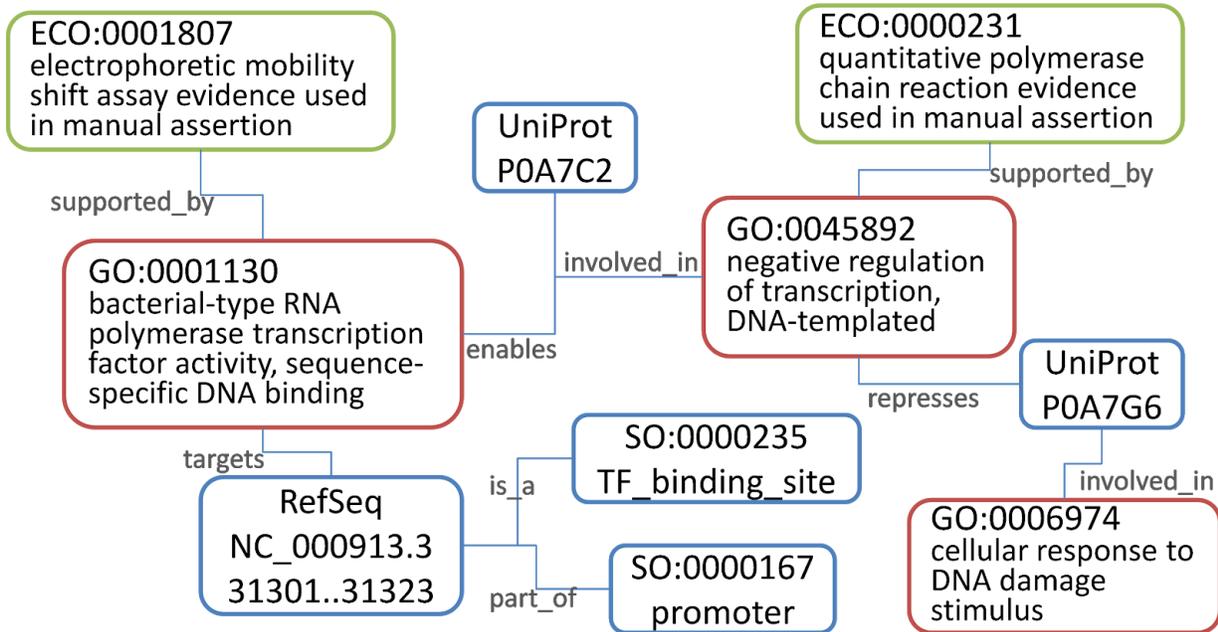


A revised scientific process workflow with direct annotation by authors.

But how do we get there? Scientists in general don't know much about how to represent knowledge, nor have much free time to get to it either. Our answer is a **semi-supervised learning system that uses an ontology as a knowledge-base**. An [ontology](#) is a formalization of knowledge in a specific domain. A pizza ontology, for instance, may define pizzas as having toppings and bases. These are called *terms* and represent "entities" in the real world. Toppings can be then defined as being of different types (cheese, pepperoni, etc.).

Starting with the crisp definition of a term in the ontology, one can then look for instances of it (references) in a corpus of unlabeled text (e.g. journal articles), and use these newly found instances to refine the current definition of a term. This process, which can be iterated many times, is called bootstrapping and it is a form of [semi-supervised learning](#).

The Erill Lab has developed a system for bootstrapping and refining ontology terms. We choose to work with a particular ontology (the [Evidence and Conclusion Ontology](#); ECO) because it is used to define and catalog experimental evidence in research articles, and hence it is transversal to all types of annotation that one wishes to perform in a research article, such as [Gene Ontology annotations](#) (the most common form of annotation performed by biocurators).



Example of an extended Gene Ontology annotation.

The only snag is that in order to be validated and refined, our system needs an *annotated corpus*. That is, we need a set of documents on which we know what ECO terms apply to each sentence (if any), so that we can thoroughly test whether our system is performing as expected or not.

For all their impressive powers, computers are way behind humans at reading text. That is why your work, as curators annotating ECO terms on a small set of documents, is so important!