The Big Mechanism Program: Changing How Science Is Done

Andrey Rzhetsky
University of Chicago, 900 East 57th Street, Chicago, IL 60637, USA

andrey.rzhetsky@uchicago.edu

Abstract

The talk will describe details of actively evolving research conducted by the UChicago consortium of the Big Mechanism program, funded by the US DARPA agency. The consortium's work focuses on: (1) probabilistic reasoning across cancer claims culled from literature which uses custom-designed ontologies; (2) the computational modelling of cancer mechanisms and pathways to automatically predict therapeutic clues; (3) automated hypothesis generation to strategically extend this knowledge, and; (4) developing a 'Robot Scientist' that performs experiments to test hypotheses probabilistically, then feeding those results back to the system.

1 Introduction

DARPA is funding the Big Mechanism program (http://www.darpa.mil/program/big-mechanism) in order to study large, explanatory models of complicated systems in which interactions have important causal effects. The program's aim is to develop technology used to read research abstracts and papers and extract pieces of causal mechanisms, assemble these pieces into more complete causal models, and reason over these models to produce explanations. The program's domain is cancer biology, with an emphasis on signalling pathways; this is just one example of causal, explanatory models, that we are hoping will be extensible across multiple domains, similar to what IBM Watson's team [1] is attempting presently.

2 The overall structure of the Big Mechanism program

The program is currently organized into three consortia, all of which take different views of causal models, different reading technologies, and different use cases.

The largest consortium, called FRIES, includes groups at CMU, SRI, University of Arizona, Oregon Health Sciences University, and others. FRIES's main focus is to explain signalling pathway behaviours. For instance, why is the expression of a gene ephemeral? Technologically, FRIES focuses on information extraction over deep reading, simulation, and even FPGA acceleration of systems biology simulators.

The second consortium ("UChicago"), in which the author of this keynote acts as the PI, is composed of researchers at the University of Chicago, the United Kingdom's National Center for Text Mining at the University of Manchester, along with participants from the Brunel University in London, all of whom collaborate on developing robotic platforms for experiment design and analysis.

The third consortium, called CURE, consists of two groups from Harvard Medical School, IHMC in Florida, and SIFT. Their focus is on deep reading, fine-grained modeling, and simulation of cell signaling's underlying biochemistry.

This talk will provide an overview of the objectives and results related mostly to the work of the second consortium.

3 UChicago consortium

As the project is ongoing and far from completion, we will cover the ideas that led the consortium to our current system design, our biological and medical motivations, and preliminary results.

Motivation: Today, cancer-related text mining is performed in linear pipelines (named entity recognition to event extraction) without explicitly estimating statement uncertainty or importance relative to a total model of cancer. Moreover, reading is divorced from reasoning and experimentation. Probabilistic reasoning is rarely used. Similarly, the Robot Scientist approach currently uses non-probabilistic logic and is disconnected from text mining and not applied to medicine. In addition, a wealth of panomics data is increasingly available, but existing methods treat each event independently and disregard prior knowledge.

Fundamental medical problem: We do not fully understand how to stop cancer cells from growing faster than normal tissue, and spreading throughout the body (metastasizing). Death from cancer typically occurs when uncontrolled growth occurs in a place where it cannot be surgically removed. Most traditional anti-cancer drugs are highly toxic to patients. As a result, single drug treatment is generally undesirable for the following reasons: (1) It is generic and not targeted to the patient and their cancer's genotype(s); (2) Intervention is

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

required at multiple points along a cancer pathway, and; (3) Cancer evolves resistance. The Holy Grail of cancer therapy is to find highly potent, non-toxic drug combinations that are tailored to individual patients, and linked to the readout of gene and protein expression from their specific cancer(s).

The system developed by the consortium incorporates three components, called Reading, Assembly, and Explanation (see Figure 1). These components integrate machine reading with probabilistic modelling, the design of custom-made ontologies, and automated experiments conducted by the Robot Scientist (a robot that is driven by experiment-designing and planning programs). For quality control and benchmarking, an independent set of experiments is conducted by humans.

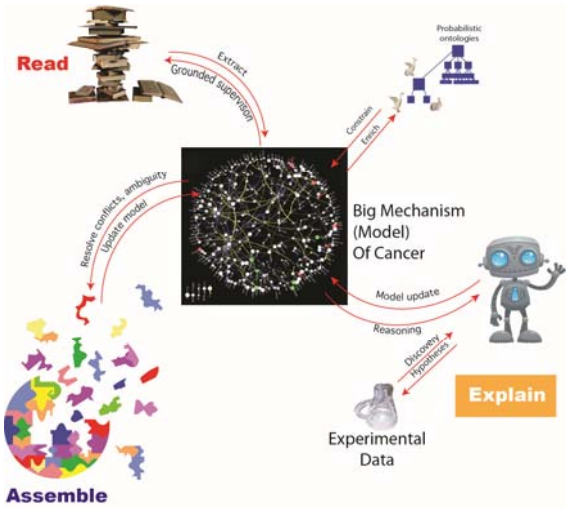

Figure 1 The integrated system, see references [2,3] for related prior work contributing to the components of the system

To illustrate how all these components come together, the talk will present a use case: Automated, optimal drug combination prediction for achieving activation or silencing of target gene(s) in a breast cancer cell line. In our initial setup, we are using a text-mined network of about three hundred genes and proteins, containing parts of networks in use cases 1 and 2. In the first pass, we focused on activating the estrogen receptor gene (ESR1) in a triple-negative breast cancer cell line by administering a cocktail of two or more FDA-approved drugs.

The motivation for the use case is to practically apply growing (through machine reading and experimental validation) model of cellular machinery to manipulate the state of the cancer cell, achieving silencing or activation of target genes/proteins in the absence of drugs specifically targeting these molecules. If successful, computationally-derived drug cocktails could at least partially reduce the need to develop new drugs, easing the economic burden of discovering and testing new medications. (Each new FDA-approved drug has an estimated price tag of somewhere between 100 million and 1 billion US dollars.)

The system generates hypotheses of the form "cocktail of drugs X1, …, Xn activates gene ESR1" and each hypothesis is tested experimentally in a triple-negative breast cancer cell line. Either human biologists or the Robot Scientist carry out these experiments.

4 "UChicago" team

Reading (NLP and text-mining; ontologies, corpus-dependent and unsupervised information extraction, logic): Sophia Ananiadou, Junichi Tsujii, Larisa Soldatova, Hoifung Poon, Andrey Rzhetsky, Robert Stevens, James Evans.

Assembling (Models of quality of science, quality of extraction, consistency, statement provenance, Markov Logic, crowdsourcing): Jacob Foster, James Evans, Hoifung Poon, Andrey Rzhetsky.

Explaining (Markov Logic, graphical models, consistency models, kinetic/dynamic consistency models): Hoifung Poon, Jacob Foster, James Evans, Ishanu Chattopadhyay, Andrey Rzhetsky.

AI and Robotics: Hoifung Poon, Kevin P. White, Ross D. King. Cancer-specific, wet-lab experiments: Ross D. King, Kevin P. White.

In prior work, Ross D. King's laboratory has developed two Robot Scientists, "Adam" and "Eve", which are among the most advanced existing laboratory automation systems.

5 Conclusion

The approach chosen by the team relies on the assimilation of massive, pre-existing literature (similar to IBM Watson) combined with iterative model updating based on empirical data and newly designed experiments (unlike IBM Watson). The project's general methodology is not domain-specific, so it is theoretically extensible across scientific domains.

References

[1] Best, J. IBM Watson: The Inside Story Of How The Jeopardy-Winning Supercomputer Was Born, And What It Wants To Do Next - Feature - TechRepublic. TechRepublic. N.p., 2015. Web. 13 May 2015.

[2] Evans J, Rzhetsky A. Machine science. Science. Jul 23 2010;329(5990):399-400.

[3] King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova, LN, Sparkes A, Whelan KE, Clare C. The Automation of Science. Science. 2010. 324, 85-89.