This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

Interactive Extreme – Scale Analytics towards Battling Cancer

Nikos Giatrakos[1,2], Nikos Katzouris[3], Antonios Deligiannakis[1,2], Alexander Artikis[4,3], Minos Garofalakis[1,2], George Paliouras[3], Holger Arndt[5], Raffaele Grasso[6], Ralf Klinkenberg[7], Miguel Ponce De Leon[8], Gian Gaetano Tartaglia[9], Alfonso Valencia[8] and Dimitrios Zissis[10,11]

[1]IMSI-Athena RIC, Greece, [2]Technical University of Crete, Greece, [3]NCSR Demokritos, Greece, [4]University of Piraeus, Greece, [5]SpringTechno, Germany, [6]NATO STO CMRE, Italy, [7]RapidMiner, Germany, [8]Barcelona Supercomputing Center, Spain, [9]Center for Genomic Regulation, Spain, [10]University of the Aegean, Greece, [11]MarineTraffic, Greece

1. INTRODUCTION

A synergetic understanding of cancer evolution and the effect of combination drug therapies on it is the cornerstone for developing effective personalized treatments, which can radically improve patients' well-being and their quality of (work and social) life. By extension, improving the treatment of patients indirectly enhances the quality of life for families, friends, and carers. Moreover, personalizing effective therapeutic approaches reduces treatment duration, cutting down healthcare monetary costs, which can be re-directed to other health and social services. Given that three out of four US families will at some point experience a family member suffering from cancer[1], the potential impact of improved cancer treatment is of considerable socio-economic and organizational significance.

Paving the way for the development of new cancer treatments requires identifying personalized drug combinations, which, by acting synergistically, will be able to fight cell resistance in target therapies and ultimately increase the patients' life expectancy and quality of life. Studying the effects of drug synergies on resistant cells is central to the problem of curing cancer. These phenomena are subject to the inherent complexity of biological systems and affected by the interplay between different processes that occur at different scales. For instance, they depend on the molecular mechanisms by which individual cells can develop resistance to a particular drug [1], which are complex in their own right, but they also depend on various types of dynamic processes concerning populations of cells. Examples of the latter include the variability in the gene expression profiles of different cells, which gives rise to heterogeneous populations, the competition for resources such as space and nutrients, as well as the interaction or cross-talk between different cells [2]. Consequently, multi-cellular systems' dynamics, such as tumor growth and evolution, can only be understood by studying how individual cells grow, divide and die, and the interactions between the cells at the population level.

In-silico models are becoming powerful tools in the fight against cancer, since they allow the combination, in a controlled environment, of heterogeneous sources of experimental data with prior biological knowledge, towards a better understanding of the underlying mechanisms and the biological processes which determine tumor growth, resistance to therapies and effects of drug synergies in cells. The multi-scale interplay of such processes makes their modeling and simulation hard. For instance, an average-sized tumor nodule contains approximately 10^8 - 10^9 cells. Each such cell needs to be modeled individually, thus its state must be updated at each step of the simulation process. Moreover, modeling the dynamics of the entire system at the cell-population level, and the evolution of various environmental parameters via partial differential equations adds additional complexity to the task. The process of simulating such a cellular system produces data at a rate of approximately 100 GB/min**.** Therefore, in-silico models of tumor evolution and resistant cells' emergence require immense computational resources, in addition to extremely efficient data processing algorithms, operating on massive data streams.

This work presents the necessary architectural and algorithmic apparatus for speeding up (a) the repetitive procedures employed by biologists in cooperation with data scientists for modelling, setting up and running such simulations, and (b) the extraction of simulation outcomes so that drug combinations are interactively applied. This way, the time for personalized and, consequently, more effective treatments is cut down.

2. TECHNOLOGICAL CHALLENGES

Below, we summarize the technological challenges that must be addressed for the development of new cancer treatments.

Challenge 1: On one hand, utilizing large computer clusters, along with High-Performance Computing (HPC) resources in order to avoid stalling clock rates when dealing with extreme-scale simulations is of utmost importance. On the other hand, more often than not, data analysts cooperating with life scientists have to resort to traditional, batch processing methodologies, where a computer cluster is treated as a black box. A user submits a job to the cluster and patiently waits for the job to finish, with little insight on what is being processed and where, what data volumes are being moved around and at what cost, how long the whole process will take and whether it could be *optimized* to run more efficiently. Often, what lies at the end of this long and expensive process, is yet another set of experiments. Consider, for instance, a life science data analyst who needs to tune a vast number of parameters for a complex machine learning model, or identify potentially useful cell or population features, with no alternative other than repeatedly running a different parameter/feature set configuration.

[1] http://natamcancer.org/NAP_Native_American_Priorities.pdf

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

Requirement 1: Addressing this challenge requires developing novel *interactive processing tools for extreme-scale data analytics*. Such tools will enable life scientists in cooperation with data analysts to iteratively pose queries and derive rapid responses. They can also provide support on how time-consuming computations may be broken down into smaller chunks in an optimal manner, based on the specifications of the underlying computational resources, and how results from such smaller chunks may be presented to the life scientist in an incremental fashion.

Challenge 2: Big Data analytics tools mine past data to extract patterns conveying insights into what has happened, and then project those patterns into the future to make sense of the fresh simulation data that stream-in. This permits only the detection of such patterns, which is often inadequate. In order to mitigate risks, capitalize on opportunities and allow for proactive decision-making, predictive analytics tools, enabling forecasts of future events of interest are required. Consider, for instance, the ability to forecast the emergence of resistant cells via the detection of small changes in physiological or molecular markers. Interactive analytics, therefore, should be guided by complex event forecasting. The ability to forecast, as early as possible, a good approximation to the outcome of a time-consuming and resource-demanding computational task also allows the rapid identification of undesired outcomes and saves valuable amount of time, effort and computational resources, which would otherwise be spent in vain. Consider, for example, the possibility to forecast the outcome of a complex multi-cellular system simulation for tumor evolution, without the need to wait for the simulation to be completed.

Requirement 2: Addressing this challenge requires cutting-edge techniques combining *distributed, online machine learning and complex event forecasting*. Algorithms should incorporate, in real-time, new knowledge that streams into ever-changing, ever-adapting, but highly robust and accurate models, while also being able to make trustworthy forecasts for future events of interest, fostering proactive decision-making. Additionally, a library of distributed, highly-scalable and optimized machine learning techniques, based on deep analysis of massive amounts of historical data, allows the discovery of valuable patterns of past behavior, which are subsequently subject to change and adaptation via online revision techniques.

Challenge 3: Allowing the data analyst to gain rapid insights on the characteristics of the massive input, lays the ground for a synergy between domain experts, human analysts, data analytics algorithms and computing infrastructures, in contrast to the current state of affairs, where algorithms and infrastructures are treated as black boxes. Such a synergy could result in efficiently handling problems, which are currently beyond the abilities of contemporary information and computer technology. As a simple example of such a synergy consider a data analyst able to pre-attentively identify an informative feature set over a multi-cellular population, using her prior domain knowledge and some first insights on the data properties, gained by exploring *data synopses*. Proposing this feature set to a machine learning algorithm may ultimately improve the algorithm's performance faster than it would be possible without human intervention. In turn, this affects the available computational resources, which can be allocated elsewhere.

Requirement 3: Interactive data analytics should foster such synergies between humans and algorithms towards the effective solution of difficult problems.

Challenge 4: Domain experts and even data analysts often do not possess the necessary programming skills to code, optimize and debug data processing operations over Big Data and HPC infrastructures.

Requirement 4: A visual tool for setting up the desired data processing operations that would automatically translate the setting to optimized code is highly desirable. In contrast to mainstream solutions, what is needed here is to design and develop a *flexible, pluggable, distributed software architecture that will be largely programmable and set up by graphical data processing workflows*. This will allow the non-data scientist, such as the life scientist, to take the most out of the underlying computational resources, by interactively querying data-at-rest and data-at-motion and experimenting with a multitude of otherwise highly-opaque models with complex parameter configurations.

3. METHOD, ARCHITECTURE & APPARATUS

Overcoming the challenges outlined in the previous section revolves around integrating two core concepts: *interactive data analytics and operational proactivity via complex event forecasting*. Figure 1 illustrates a possible architecture and outlines an algorithmic apparatus which unites the two concepts.

As shown in the figure, extreme-scale data streams are continuously acquired from the various (cellular system simulation) sources. A dedicated, lightweight engine relying on approximate query processing techniques is introduced to extract compact synopses out of these data streams [3][4][5]. Synopses are thought of as a prerequisite that allows for fast response times during exploring the massive, high-speed input (in Figure 1 this corresponds to Synopses Data Engine - SDE). Then, both the extracted data synopses and data streams are forwarded to a number of recommended components responsible for interactive analytics and operational proactivity. Accounting for the requirements in Section 2 involves four main components for: (a) distributed machine learning & data mining; (b) data processing and workflow specification; (c) distributed complex event forecasting and (d) optimization and runtime adaptation of Big Data processing workflows. We examine the operation of each component in turn, below.

The distributed machine learning & data mining component is used to incorporate a wide variety of tools for building various types of predictive models from data. These tools can rely on existing algorithms for data analysis [6][7], which need to be customized to work in a distributed, streaming setting, and optimized to cope with the immense data volumes. The machine learning algorithms are utilized to constantly mine the streaming input to generate, maintain and update a rich variety of diverse, high-quality models and model ensembles. These models can be used, at any time, to analyze the data that stream-in. At the same time, these algorithms may be used as a library for interactive data analytics, for the data analyst to experiment with compressed data synopses. This will enable the data analyst to e.g. try different parameter configurations per model, different feature sets on existing models, or even generate new models on-the-fly from

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

representative data views, thus gaining early insights on the properties of the data she's dealing with.

The process of interactive data analytics is not supposed to hinder the actual knowledge extraction from the streaming data: for instance, the life scientist should be able to explore data stream synopses, as the machine learning algorithms seamlessly integrate new regularities into the models, by continuously updating them from the entire stream of data.

The data processing and workflow specification component is added to offer additional support for interactive data analytics, via a rich variety of data processing functions, as well as the ability to plug-in new custom functions, or new data analytics algorithms (see Figure 1). Moreover, it allows the composition of data analytics pipelines from existing models and data processing functions. Interactive data analytics must be supported by graphical interfaces for the specification of data processing workflows, thus enabling the domain expert, in cooperation with the data analyst, to take the most out of the platform with minimum coding effort. The vision is to involve non-data scientists into fast loops of interactive, exploratory data analysis, where domain experts and data analysts may gain rapid insights on properties of the data "at a glance".

The optimization and runtime adaptation of query execution is a component suggested to facilitate fast response times during extreme-scale analytics across different computing platforms, as well as the execution for Big Data processing workflows (see Figure 1). Data analytics is not necessarily performed on a single platform. Parts of the processing could be pushed to the input sensor level e.g., collecting simulation data, while more computationally intensive operations, such as population cell simulations, could be executed in one or more (potentially distributed) Big Data platforms, or within clusters (e.g. GPUs) of a supercomputer. Even within a single supercomputer, one often finds different available clusters, with different hardware and processing capabilities. In case of (machine learning) operators that have available implementations in different platforms, it is desirable to minimize the data analysts' involvement in the specification of the platform on which these operators must be executed. Hence, optimization and runtime adaptation technology [8] is foreseen to automate the optimal allocation of resources on-the-fly for the execution of these operations.

Distributed complex event forecasting adds up as a key enabler component for the recommended architecture. The streaming input is constantly matched against a set of event patterns, i.e. arbitrarily complex combinations of time-stamped pieces of information. An event pattern can either be fully matched against the streaming data, in which case events are *detected*, or partially matched, in which case events are *forecast* with various degrees of certainty. The latter usually stems from stochastic models of future behavior [9], embedded into the event processing loop, which estimate the likelihood of a full match, i.e. the actual occurrence of a particular complex event.

Notably, "*forecasting*" in this context is not to be confused with "*predicting*", typically used to refer to machine learning models classifying previously unseen instances. From a methodological standpoint, it also differs from time-series forecasting. Complex event forecasting combines symbolic and numerical streams for foreseeing the occurrence of any type of situation that may be defined as an event, based on combinations of other similar events and contextual knowledge.

Given that the input consists of a multitude of data streams, events may correlate sub-events across many different streams, with different attributes and time granularities. Therefore, a highly-expressive event pattern specification language, capable of capturing complex relations between events, is necessary. Moreover, the actual patterns of what constitutes an event of interest are often not known in advance, and even if they are, event patterns need to be frequently updated to cope with the drifting nature of streaming data. Therefore, the required algorithmic apparatus must incorporate machine learning techniques for learning and revising complex event patterns from data, which is another role of the component of distributed learning of event patterns (see Figure 1).

What is necessary here is highly-expressive, declarative event pattern specification formalisms, which combine first-order logic, probability theory [10][11] and automata theory [12][13]. This is due to the fact that such formalisms have a number of key advantages: (i) they are capable of expressing arbitrarily complex relations and constraints between events; (ii) they can be used for event forecasting, offering support for robust temporal reasoning; (iii) they offer direct connections to machine learning techniques for refining event patterns, or learning them from scratch, via tools and methods from the field of Statistical Relational Learning [14]. Notably, all existing techniques for complex event detection and forecasting, as well as machine learning techniques for the automatic extraction of event patterns from data, need to be extended to work in a highly-distributed, streaming setting.

Complex event forecasting contributes significantly to the effectiveness of interactive analytics. The idea is to use an event-based methodology to model time-consuming, computationally-demanding operations, and then use event forecasting to derive good approximations of the operations' outcomes, without the need to wait for the actual operations to terminate. This will save time, effort and resources (computing power, bandwidth), by enabling data analysts get rapid responses from expensive operations, thus identifying undesired outcomes and poorly performing configurations of particular data analysis approaches and facilitating the exploration of more promising options. Equally beneficial in speeding-up time-consuming computational processes to foster interactivity, is to use specialized domain knowledge for the problem at hand, which allows simplification of the computational process, by e.g. pruning large parts of a search space for some difficult optimization task, which are known beforehand to be pointless to search.

4. A VIRTUAL LABORATORY FOR STUDYING TUMOR GROWTH AND EVOLUTION

The introduced extreme-scale analytics architecture is a key enabler to provide a "virtual laboratory" for studying tumor growth and evolution. It can support the goal of using *in-silico* models of cell systems found in *in-vivo* tumors, to facilitate the design, testing and optimization of cancer treatments based on combinations of different drugs.

In-silico simulations of such phenomena require modeling multi-scale processes, thus bridging the gap between different levels of

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

description and connect events that occur at different scales. As an example of such correlated multi-scale phenomena, consider a DNA mutation that alters the function of a protein, which leads to an alteration of the cell-cycle — the process by which a cell replicates its DNA and divides into two daughter cells — producing an uncontrolled cell growth [1]. At the molecular level, an individual cell can be described by the network of its signal transduction pathways [15][16]. Modeling this network allows the representation of the molecular machinery by which a cell integrates environmental signals and alters the gene expression patterns in response to specific stimuli. For example, when a growth factor binds to a membrane receptor (the stimulus), this signal is translated into an activation of the cell cycle, i.e., the cell starts to grow in order to divide into two daughter cells. Models of signaling networks have been reconstructed for different cell types and can be simulated using different approaches, such as Ordinary Differential Equation or the Boolean Formalism [16][17]. Such models describe the state of different signaling pathways, which can be used to predict the "fate" of a particular cell (proliferation/apoptosis). In turn, these models can be used to predict the effects of drug synergies, as in the case of [15], where a model reconstructed and calibrated for the gastric adenocarcinoma cell-line (AGS) was used to accurately predict (as validated through growth experiments) the effects of drug synergies.

At the cell-population level, a simple Agent-Based Model (see Figure 2c) has been developed that takes into account gene expression levels (see Figure 2a), growth, as well as nutrient consumption, and where each individual agent has its own signaling network from which the propensity, for an individual cell, to proliferate or to enter into apoptosis can be calculated using Boolean simulations (see Figure 2b). This model is being used to integrate and interpret experimental data on gene variability and drug synergies. A fundamental feature of simulating cell populations is that it allows for modeling heterogeneity. This is of importance in the study of cancer, since tumor heterogeneity and resistance to target therapies are two closely interconnected phenomena [18]. Integrating experimental information of gene expression variability available from RNA sequence data as well as DNA mutations (see Figure 2a), it is possible to generate heterogeneous populations of agents (see Figure 2c) which resemble the variability known to be present in tumors [17].

Simulation of multi-cellular systems is computationally demanding, but because of the simulation structure, it is very suitable for running in HPC environments. Since, at each time step the internal state of each agent can be computed independently of the other agents, these operations can be run in parallel and thus the performance may scale linearly with the number of agents. There are already different packages for simulating multi-scale models, and some of them have been developed to run in HPC environments (see Figure 2d). For example, PhysiCell has already been designed and implemented for HPC environment and its deployment is straightforward [19]. However, simulating the evolution of multi-cellular systems of realistic sizes, involving billions of cells, as those found in in-vivo tumors remains a challenging task, even for HPC infrastructures. Moreover, to study the effects of drug synergies on such systems requires a very large number of such simulations.

To support the simulation of multi-cellular systems with the actual number of cells found in in-vivo tumors, the framework outlined in Section 3 would host, in the respective components, (i) machine and deep learning techniques for obtaining dynamic cell-cycle models; (ii) complex event forecasting techniques for the early detection of undesired simulation outcomes, e.g. when performing large-scale parameter screening, as well as various events of biological interest, such as forecasting the emergence of resistant cells via the detection of small changes in physiological or molecular markers; (iii) graphical workflow designs and interactive learning techniques enhancing the efficiency of model calibration and parameter selection during the simulation process. All the above would be efficiently orchestrated by a optimization and runtime adaptation component.

Such a virtual laboratory will lead to the exploitation of simulation outcomes for the timely indication of personalized therapies. Socio-economic implications of improved therapies, in turn, will aid patients to maintain their working status and quality of life as well as receive equal treatment from insurance, banking (such as loan) or other services in their everyday life [20]. Moreover, shorter treatment duration will cut down public healthcare costs, with the possibility of savings being invested to other social services.

5. BROADER APPLICATION DOMAINS

In principle, the generic architecture of Figure 1 suits several other application domains dealing with massive data flows. Taking advantage of such data requires sophisticated analytics tools, capable of extracting insights on-the-fly, from a multitude of voluminous, correlated, high-velocity data streams, but also, harvesting ever-growing historical data repositories. This imposes similar, compared to life sciences, challenges to computer scientists, system engineers and industrial stakeholders.

Maritime Monitoring: Maritime surveillance applications need to combine high-velocity data streams, including global maritime surveillance systems, such as the AIS (Automatic Identification System), with acoustic signals of autonomous, unmanned vehicles, such as Wave gliders [21]. Employing an architecture which incorporates the discussed algorithmic apparatus helps improving maritime situational awareness by enabling the accurate identification (via learning new patterns) and forecasting the activities of "dark targets" that are (intentionally) hidden from traditional monitoring systems.

Finance: In the financial domain, stock price forecasting and portfolio management rely on stock tick data combined with rich, real-time information sources on various pricing indicators. Financial data involve a variety of market data, including stock market and crypto-currencies market data, arriving in tens of thousands of correlated, high-velocity streams. The distributed complex event forecasting component of Figure 1 can be used to forecast price swings of stocks, currencies, commodities and the distributed learning and data mining component can serve systemic risk prediction purposes and offer decision support for investment opportunities.

6. SUMMARY

We presented the method, architecture and algorithmic apparatus for materializing interactive extreme-scale analytics in the battle against cancer. Our discussion expands on the whole processing

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

pipeline, from the time distributed data streams from simulations of multi-scale biological processes are digested into a Big Data/HPC platform, to the extraction of real-time knowledge and event forecasting. The apparatus includes architectural components equipped with the novel algorithms necessary for constructing concise data summaries, machine learning models for knowledge extraction and event forecasting facilities. These components are assisted by integrated workflow design tools, minimizing programming effort, and by an optimizer transparently devising the execution of the whole data processing pipeline. This way, the time to set up on-line processing pipelines is diminished, and the real-time discovery and monitoring of events is enabled, paving the way for searching, discovering and employing new personalized cancer therapies.

7. ACKNOWLEDGMENTS

This work is supported by the European Commission under the INFORE project (H2020-ICT- 825070).

8. REFERENCES

[1] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[2] S. M. Shaffer *et al.*, “Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance,” *Nature*, vol. 546, no. 7658, pp. 431–435, Jun. 2017.

[3] M. Garofalakis, J. Gehrke, R Rastogi (Eds.). "Data Stream Management -- Processing High-Speed Data Streams", Springer-Verlag, New York (Data-Centric Systems and Applications Series), July 2016.

[4] F. Li, B. Wu, K. Yi, and Z. Zhao, “Wander Join: Online Aggregation via Random Walks,” in Proceedings of the 2016 International Conference on Management of Data, New York, NY, USA, 2016, pp. 615–629.

[5] N. Giatrakos, A. Deligiannakis, M. Garofalakis, D. Keren, V. Samoladas, “Scalable approximate query tracking over highly distributed data streams with tunable accuracy guarantees”, Information Systems. 1;76:59-87, July 2018.

[6] N. Katzouris, A. Artikis, and G. Paliouras, “Incremental learning of event definitions with Inductive Logic Programming,” Mach. Learn., vol. 100, no. 2–3, pp. 555–585, Sep. 2015.

[7] N. Katzouris, A. Artikis, and G. Paliouras, “Online learning of event definitions,” TPLP, vol. 16, no. 5–6, pp. 817–833, 2016.

[8] I. Flouris et al, “FERARI: A Prototype for Complex Event Processing over Streaming Multi-cloud Platforms”, SIGMOD 2016.

[9] E. Alevizos, A. Artikis, and G. Paliouras, “Event Forecasting with Pattern Markov Chains,” DEBS 2017.

[10] A. Artikis, M. J. Sergot, G. Paliouras: An Event Calculus for Event Recognition. IEEE TKDE. 27(4): 895-908 (2015)

[11] A. Skarlatidis et al. A Probabilistic Logic Programming Event Calculus. TPLP, 2014.

[12] I. Flouris, N. Giatrakos, A. Deligiannakis, M. Garofalakis, M. Kamp, M. Mock, “Issues in complex event processing: Status and prospects in the Big Data era”, Journal of Systems and Software 127: 217-236 (2017)

[13] N. Giatrakos, A. Artikis, A. Deligiannakis, M. Garofalakis, “Complex Event Recognition in the Big Data Era”, PVLDB 10(12): 1996-1999 (2017)

[14] L. D. Raedt, K. Kersting, S. Natarajan, and D. Poole, “Statistical Relational Artificial Intelligence: Logic, Probability, and Computation,” Synth. Lect. Artif. Intell. Mach. Learn., vol. 10, no. 2, pp. 1–189, Mar. 2016.

[15] L. Calzone *et al.*, “Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement,” *PLOS Comput. Biol.*, vol. 6, no. 3, p. e1000702, Mar. 2010.

[16] P. Bloomingdale, V. A. Nguyen, J. Niu, and D. E. Mager, “Boolean network modeling in systems pharmacology,” *J. Pharmacokinet. Pharmacodyn.*, vol. 45, no. 1, pp. 159–180, Feb. 2018.

[17] E. Kim, J.-Y. Kim, M. A. Smith, E. B. Haura, and A. R. A. Anderson, “Cell signaling heterogeneity is modulated by both cell-intrinsic and -extrinsic mechanisms: An integrated approach to understanding targeted therapy,” *PLOS Biol.*, vol. 16, no. 3, p. e2002930, Mar. 2018.

[18] I. Dagogo-Jack and A. T. Shaw, “Tumour heterogeneity and resistance to cancer therapies,” *Nat. Rev. Clin. Oncol.*, vol. 15, no. 2, pp. 81–94, Feb. 2018.

[19] A. Ghaffarizadeh, R. Heiland, S. H. Friedman, S. M. Mumenthaler, and P. Macklin, “PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems,” *PLOS Comput. Biol.*, vol. 14, no. 2, p. e1005991, Feb. 2018.

[20] F. Mols, et al. "Socio-economic implications of cancer survivorship: results from the PROFILES registry." European Journal of Cancer 48.13 (2012): 2037-2042.

[21] A. Tesei, S. Fioravanti, V. Grandi, P. Guerrini, and A. Maguer, “Localization of small surface vessels through acoustic data fusion of two tetrahedral arrays of hydrophones,” *Proc. Meet. Acoust.*, vol. 17, no. 1, p. 070050, Jul. 2012.

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

AUTHOR BIOS

Holger Arndt (h.arndt@springtechno.com) is Managing Director at Spring Techno with 10+ years of experience in international R&D projects.

Alexander Artikis (a.artikis@unipi.gr) is an Assistant Professor in the Department of Maritime Studies of the University of Piraeus, and a Research Associate in the Institute of Informatics & Telecommunications at NCSR "Demokritos", in Athens, Greece, where he leads the Complex Event Recognition lab.

Antonios Deligiannakis (adeli@softnet.tuc.gr) is an Adjunct Researcher at ATHENA-IMSI and an Associate Professor of Computer Science at the School of Electrical and Computer Engineering of the Technical University of Crete. He is also the coordinator of the INFORE project.

Minos Garofalakis (minos@softnet.tuc.gr) is the Director of ATHENA-IMSI and a Professor at the Technical University of Crete since 2008.

Nikos Giatrakos (ngiatrakos@softnet.tuc.gr) is an Adjunct Researcher at ATHENA-IMSI and a Postdoctoral Research Associate at the School of Electrical and Computer Engineering of the Technical University of Crete.

Raffaele Grasso (raffaele.grasso@cmre.nato.int) is a scientist in the research department of the NATO Science & Technology Organization-Centre for Maritime Research and Experimentation (CMRE).

Nikos Katzouris (nkatz@iit.demokritos.gr) is a Research Associate at the the Institute of Informatics & Telecommunications, at NCSR "Demokritos", in Athens, Greece.

Ralf Klinkenberg (rklinkenberg@rapidminer.com) is a Co-Founder of RapidMiner and Head of Data Science Research, since 10/2007.

George Paliouras (paliourg@iit.demokritos.gr) is a Senior Researcher and head of the Intelligent Information Systems division of the Institute of Informatics and Telecommunications at NCSR "Demokritos", in Athens, Greece.

Miguel Ponce de León (miguel.ponce@bsc.es) joins Barcelona Supercomputer Center as a Postdoctoral Researcher at the Computational Biology Group of the Life Science Department. He is also a scientist at the National System of Researchers (SNI, Uruguay).

Gian Gaetano Tartaglia (gian@tartaglialab.com) Since 2010 Gian is a PI at the Centre for Genomic Regulation (CRG) in Barcelona. Since 2014, Gian is tenured in Catalonia as ICREA professor of Life and Medical Sciences.

Alfonso Valencia (alfonso.valencia@bsc.es) is an ICREA Professor and a senior scientist with activity in various areas of Bioinformatics and Computational Biology.

Dimitrios Zissis (dzissis@marinetraffic.com) is an Associate Professor at the Department of Product & Systems Design Engineering at the University of the Aegean and Head of Research at MarineTraffic.

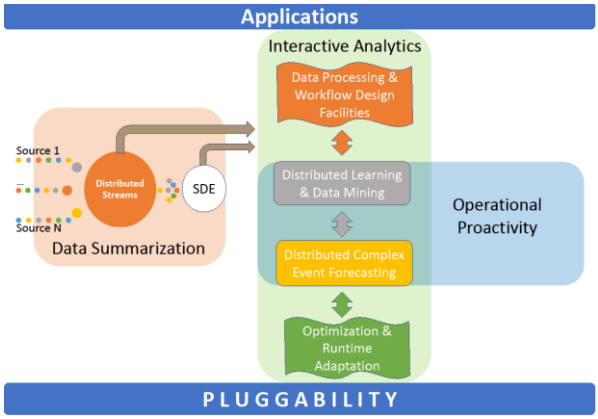

Figure 1: Interactive Extreme-scale Analytics – The Big Picture. "Applications" emphasize the applicability of the approach to several domains, while "Pluggability" refers to component modularity.

This article has been published in the IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2019.2913071

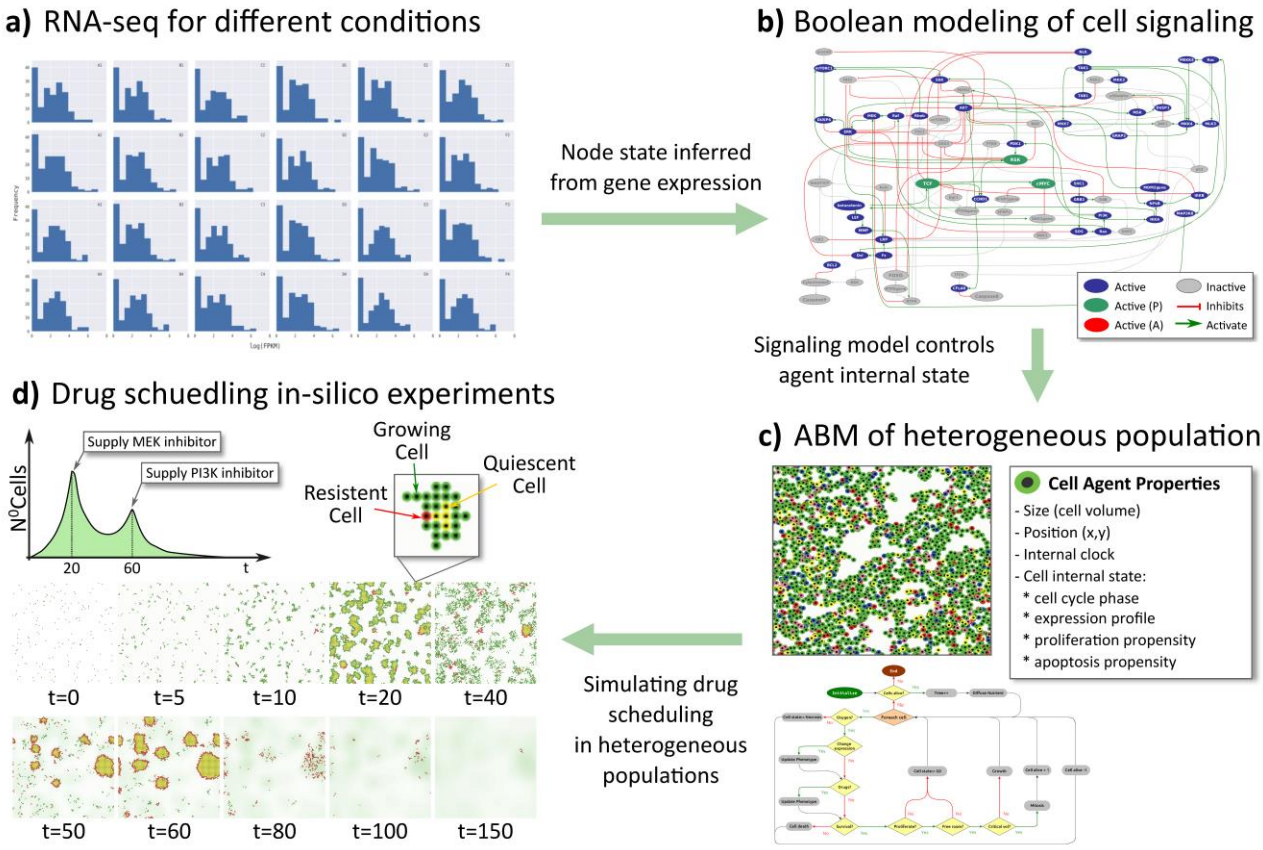

Figure 2: The simulation process of the hybrid multi-scale model of individual cell signaling. a) RNA-seq expression profiles for different conditions (i.e. different drugs, controls) for each node of the model. b) Signaling model calibrated for the gastric adenocarcinoma (AGS) cell line. Nodes are set to active or inactive based on the RNA-seq data. c) Different signaling networks are embedded into an Agent-Based Model (ABM) to simulate population level dynamics. d) In-silico experiments of cell cultures treated with different drugs.