Citizen Data Science for Social Good in Complex Systems: Case Studies and Vignettes from Recent Projects

Soumya Banerjee [1,2]
[1] University of Oxford
Oxford, United Kingdom
[2] Ronin Institute
Montclair, USA

soumya.banerjee@maths.ox.ac.uk

Introduction

The confluence of massive amounts of openly available data, sophisticated machine learning algorithms and an enlightened citizenry willing to engage in data science presents novel opportunities for crowd sourced data science for social good. In this submission, I present vignettes of data science projects that I have been involved in and which have impact in various spheres of life and on social good.
Complex systems are all around us: from social networks to transportation systems, cities, economies and financial markets. Understanding these complex systems may lead to solutions for problems ranging from famines, global crises, poverty, climate change and sustainable living despite over-population. Big data and citizen data science allows unprecedented computational power and collective intelligence to be brought to bear on fundamental challenges facing humanity like poverty, diseases, famines and developmental challenges.

Crime in societies

Using openly available data from the US Census and FBI combined with machine learning techniques, we uncover novel patterns of crime in US cities [1,2]. Our results have implications for public policy especially the number of police that should be allocated in larger cities and budget for law enforcement.

We look at freely available data about violence and assault on women in US college campuses. Using machine learning techniques we uncover trends and patterns that highlight the need for protection of women and greater transparency in how universities handle cases of assault [4]. We have also built and freely shared tools that allow people to interact with the code and data [5]. These tools have the dual purpose of achieving crowdsourced citizen data science as well as outreach and engagement, thereby spreading awareness of relevant social issues.

Public health and emerging diseases

Global pandemics are on the rise. Novel disease like Zika and Ebola virus jump from species to species and ultimately affect humans. Using data from the Center for Disease Control (for West Nile virus) coupled with advanced machine learning techniques, we predict species that

may likely be infected in the next pandemic [3]. We also predict infectivity of viruses from very sparse experimental data [4]. These kinds of techniques can help rapidly predict the potential of emerging viruses to spread, especially when we have very little experimental data about them.

In rare cases, the immune system can attack the cells of the host organism causing autoimmune diseases. We implemented a computational framework that combines bioinformatics and network analysis with an emerging targets platform [5]. The computational framework can be used to find drug targets for autoimmune diseases. It can also be used to find existing drugs that can be repurposed to treat autoimmune diseases based on networks of interactions or similarities between different diseases. Our computational framework uses open data on drug targets to find novel therapeutics for autoimmune diseases and potentially even other dysfunctions.

The code and associated material is available online [6]. An open source framework enables anyone with a computer and an internet connection to start searching for drug targets. Such kinds of frameworks can enable citizen scientists to contribute to drug science.

Society and developing nations

Scientific collaboration networks are an important component of scientific output and contribute significantly to expanding our knowledge and to the economy and gross domestic product of nations. We examined data from the Mendeley scientific collaboration network. We analyzed this data using a combination of machine learning techniques and dynamical models [7]. We highlight inequalities in global networks of scientific collaboration. This has implications for how developing nations invest in science and are able to make economic progress. Our model and analysis gives insights and guidelines into how scientific development of developing countries can be guided. This is intimately related to fostering economic development of impoverished nations and creating a richer and more prosperous society.

Citizen data science for complex systems

Complex systems are all around us: from social networks to transportation systems, cities, economies and financial markets. Understanding these complex systems may lead to solutions for problems ranging from famines, global crises, poverty, climate change and sustainable living despite over-population. Understanding complex systems and solving real world problems will need building multi-scale computational models that integrate understanding from multiple levels of aggregation. Such computational models will have to be 1) scalable, 2) need to make inferences from huge amounts of data (big data) and 3) practitioners will have to talk with different stakeholders to understand problems and communicate solutions to them. Computational models that scale will be critical in understanding complex systems: disease models, socio-economic systems, biological systems.

We have made all our code and analysis available online [8] (http://doi.org/10.5281/zenodo.883783). We hope this will allow any citizen scientist to engage in model building and hypothesis testing.

We need citizen scientists enabled with open data and freely available computational techniques to engage with humanity's pressing problems. Big data and citizen data science allows unprecedented computational power and collective intelligence to be brought to bear on fundamental challenges facing humanity like poverty, diseases, famines and developmental challenges.

References

[1] Soumya Banerjee, Pascal van Hentenryck, Manuel Cebrian. Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities, Palgrave Communications, 2015

[2] Soumya Banerjee. An Immune System Inspired Theory for Crime and Violence in Cities. Interdisciplinary Description of Complex Systems, 15(2):133-143, 2017

[3] Soumya Banerjee. Scaling in the Immune System, PhD Thesis, University of New Mexico, USA, 2013

[4] Soumya Banerjee, Jeremie Guedj, Ruy Ribeiro, Melanie Moses, Alan Perelson (2016). Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-. http://doi.org/10.1098/rsif.2016.0130

[5] Soumya Banerjee. (2017) A bioinformatics and network analysis framework to find novel therapeutics for autoimmunity, PeerJ Preprints 5:e3217v2 https://doi.org/10.7287/peerj.preprints.3217v2

[6] Soumya Banerjee. (2017). A bioinformatics and network analysis framework to find novel therapeutics for autoimmunity: Supplementary Resources. Zenodo. http://doi.org/10.5281/zenodo.883777

[7] Soumya Banerjee, Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns, arXiv preprint arXiv:1509.07313, 2015

[8] Soumya Banerjee. (2017). Citizen Data Science for Social Good: Case Studies and Vignettes from Recent Projects (Supplementary Resources). Zenodo. http://doi.org/10.5281/zenodo.883783

[9] https://bitbucket.org/neelsoumya/public_open_source_datascience, URL accessed Sept 2017.

[10] A Biologically Inspired Model of Distributed Online Communication Supporting Efficient Search and Diffusion of Innovation, Soumya Banerjee, Interdisciplinary Description of Complex Systems 14 (1), 10-22, 2016

[11] A spatial model of the efficiency of T cell search in the influenza-infected lung L Drew, F Stephanie, B Soumya, C Candice, C Judy, M Melanie. Journal of Theoretical Biology 398 (7), 52-63

[12] Banerjee, S., Guedj, J., Ribeiro, R. M., Moses, M., & Perelson, A. S. 2016. Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-.http://doi.org/10.1098/rsif.2016.0130

[13] Science and technology consortia in US biomedical research: A paradigm shift in response to unsustainable academic growth. Curt Balch, Hugo Arias-Pulido, Soumya Banerjee, Alex K. Lancaster. BioEssays 37 (2), 119-122

[14] Soumya Banerjee and Joshua Hecker. A Multi-Agent System Approach to Load-Balancing and Resource Allocation for Distributed Computing, arXiv preprint arXiv:1509.06420, 2015

[15] Soumya Banerjee and Melanie Moses. Immune System Inspired Strategies for Distributed Systems. arXiv preprint arXiv:1008.2799, 2010

[16] Soumya Banerjee and Melanie Moses. Scale Invariance of Immune System Response Rates and Times: Perspectives on Immune System Architecture and Implications for Artificial Immune Systems. Swarm Intelligence 4, 301–318 (2010). URL http://www.springerlink.com/content/w67714j724486331/

[17] Soumya Banerjee, A Roadmap for a Computational Theory of the Value of Information in Origin of Life Questions, Interdisciplinary Description of Complex Systems, 14(3), 314-321, 2016

[18] Soumya Banerjee, Jeremie Guedj, Ruy Ribeiro, Melanie Moses, Alan Perelson (2016). Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-. http://doi.org/10.1098/rsif.2016.0130

[19] Soumya Banerjee. 2009. An Immune System Inspired Approach to Automated Program Verification, arXiv preprint arXiv:0905.2649, 2009

[20] Soumya Banerjee. 2013. Scaling in the immune system, PhD Thesis, University of New Mexico (2013)

[21] Soumya Banerjee. A Biologically Inspired Model of Distributed Online Communication Supporting Efficient Search and Diffusion of Innovation. Interdisciplinary Description of Complex Systems 14 (1), 10-22, 2016

[22] Soumya Banerjee. A computational technique to estimate within-host productively infected cell lifetimes in emerging viral infections. PeerJ Preprints 4 (e2621v2) 2017

[23] Soumya Banerjee. An artificial immune system approach to automated program verification: Towards a theory of undecidability in biological computing. PeerJ Preprints 5 (e2690v1) 2017

[24] Soumya Banerjee. An artificial immune system approach to automated program verification: Towards a theory of undecidability in biological computing. PeerJ Preprints 5 (e2690v1) 2017

[25] Soumya Banerjee. An Immune System Inspired Theory for Crime and Violence in Cities. Interdisciplinary Description of Complex Systems, 15(2):133-143, 2017

[26] Soumya Banerjee. Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns. arXiv preprint arXiv:1509.07313, 2015

[27] Soumya Banerjee. Optimal strategies for virus propagation. arXiv preprint arXiv:1512.00844, 2015