

CITIZEN DATA SCIENCE FOR SOCIAL GOOD IN COMPLEX SYSTEMS

Soumya Banerjee*

University of Oxford
Oxford, United Kingdom

Ronin Institute
Montclair, United States of America

DOI: 10.7906/16.1.6
Regular article

Received: 8 October 2017.
Accepted: 17 January 2018.

ABSTRACT

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.

KEYWORDS

citizen data, Zenodo, complex systems

CLASSIFICATION

JEL: C51, C88

*Corresponding author, η: soumya.banerjee@ronininstitute.org; +1 505 277 3122;
Department of Computer Science, 1, University of New Mexico, Albuquerque, NM, 87131, USA

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 [3]. We have also built and freely shared tools that allow people to interact with the code and data [4]. 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 [5]. We also predict infectivity of viruses from very sparse experimental data [3]. 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 [4]. 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.

Decentralized non-institutional collaborative networks like the Ronin Institute will enable greater citizen involvement and democratize science [8]. Decentralized collaboration networks can accelerate scientific discovery [9]. Such initiatives have the potential to engage new scientists in solving the problems of humanity despite the funding problems in science [10].

All the accompanied code and analysis are available online [11]. 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.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Alex Lancaster and Dr. Jon Wilkins for fruitful discussions.

REFERENCES

- [1] Banerjee, S.; van Hentenryck, P. and Cebrian, M.: *Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities*. Palgrave Communications **1**(15022), 2015, <http://dx.doi.org/10.1057/palcomms.2015.22>,
- [2] Banerjee, S.: *An Immune System Inspired Theory for Crime and Violence in Cities*. Interdisciplinary Description of Complex Systems **15**(2), 133-143, 2017, <http://dx.doi.org/10.7906/indecs.15.2.2>,

- [3] Banerjee, S. et al.: *Estimating Biologically Relevant Parameters under Uncertainty for Experimental Within-Host Murine West Nile Virus Infection*.
Journal of the Royal Society Interface **13**(117), 2016,
<http://dx.doi.org/10.1098/rsif.2016.0130>,
- [4] Banerjee, S.: *A bioinformatics and network analysis framework to find novel therapeutics for autoimmunity*.
PeerJ Preprints 5:e3217v2, 2017,
<http://dx.doi.org/10.7287/peerj.preprints.3217v2>,
- [5] Banerjee, S.: *Scaling in the Immune System*. Ph.D. Thesis.
University of New Mexico, Albuquerque, 2013,
- [6] Banerjee, S.: *A bioinformatics and network analysis framework to find novel therapeutics for autoimmunity: Supplementary Resources*.
Zenodo, 2017,
<http://dx.doi.org/10.5281/zenodo.883777>,
- [7] Banerjee, S.: *Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns*.
preprint arXiv:1509.07313 [cs.SI], 2015,
- [8] -: *Ronin Institute*.
<http://ronininstitute.org>, accessed 2nd November 2017,
- [9] Banerjee, S.: *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,
<http://dx.doi.org/10.7906/indecs.14.1.2>,
- [10] Balch, C.; Arias-Pulido, H.; Banerjee, S.; Lancaster, A.K.: *Science and technology consortia in US biomedical research: A paradigm shift in response to unsustainable academic growth*.
BioEssays **37** (2), 119-122,
<http://dx.doi.org/10.1002/bies.201400167>,
- [11] Banerjee, S.: *Citizen Data Science for Social Good: Case Studies and Vignettes from Recent Projects (Supplementary Resources)*.
Zenodo, 2017,
<http://dx.doi.org/10.5281/zenodo.883783>.