Data Science for Earth: An Earth Day Report

Emre Eftelioglu
Amazon
Bellevue, WA
efteli@amazon.com
affiliated with Cargill Inc. at the time of the event

Shashi Shekhar
University of Minnesota
Minneapolis, MN
shekhar@umn.edu

James Hudson
AI for Good Foundation
New York, NY
james.hodson@ijs.si

Lucas Joppa
Microsoft Research
Redmond, WA
lujoppa@microsoft.com

Chaitanya Baru
University of California
San Diego, CA
chaitanya
baru@sdsc.edu

Vandana Janeja
University of Maryland
Baltimore, MD
vjaneja@umbc.com

ABSTRACT

At the 25th ACM SIGKDD conference on Knowledge and Data Discovery (KDD), a special Earth Day symposium was held to bring together thought leaders in academia, industry and government to discuss opportunities for data science to help meet the challenges facing our planet Earth. The all-day event showcased some examples of data-intensive research being done to study Earth-related phenomena. A key goal of the symposium was to raise awareness about the unique challenges that Earth-related datasets pose for data mining. The Earth Day website can be found at https://www.kdd.org/kdd2019/special-days/earth-day.

Keywords

Earth, Sustainability, Climate Change, Urbanization, Food-Energy-Water Nexus, Geospatial Data Science.

1. INTRODUCTION

Our planet and civilization are facing major challenges from climate change and environmental degradation. Extreme events are becoming more extreme and frequent. Surface water has more pollution and greenhouse gases have increased in the atmosphere. The largest freshwater sources, polar ice-caps and glaciers, are losing ice, leading to rising sea-levels.

Knowledge discovery and data mining (KDD) is crucial to address these challenges our changing planet is facing. Earth data (e.g. geo-referenced data, in-situ and remotely sensed Earth observations, census data, gps trajectories, etc.) helps us understand biological, physical, and social changes. It can help forecast rates of sea level change in polar ice shelves and predict critical atmosphere and geospace events. It is also important for many societal priorities including public health, urbanization, transportation, and food, energy and water security and sustainability. At the same time, Earth data has unique characteristics that bring challenges to data science because of its often spatial and/or physical nature.

The goals of the Earth Day [2] were to: (1) Bring together thought leaders in academia, industry and government to explore the challenges that Earth faces today and discuss opportunities and responsibilities of the data science community. (2) Create an awareness about these challenges in the KDD community. (3) Innovate and leverage new and existing data science methods to tackle the grand challenge of climate change from a different angle.

2. WORKSHOPS

Morning sessions of the Earth Day were organized under three half-day workshops. The workshops were selected to discuss the role data science is playing to address the United Nations Sustainable Development Goals for 2030 [14]. Leaders from 193 countries identified 17 goals to provide a better future for humanity, including addressing poverty and hunger, and providing safety from the worst effects of climate change. Out of these 17 goals, the workshops investigated 8. Figure 1 shows which goals each workshop considered. The workshops had scientific paper presentations, panel discussions, as well as poster sessions (Table 1) to bring the KDD community's attention to the overarching themes of Earth Day.

2.1 Fragile Earth: Theory Guided Data Science to Enhance Scientific Discovery

The Fragile Earth workshop brought research, industry, and policy community members together to discuss how to enhance scientific discovery in the Earth sciences through the joint use of data, theory, and computation. During two research paper presentations, two keynotes and a panel discussion, participants considered following topics: "Paradigms for enhancing scientific discovery through theory guided data science", "Empirical investigations at the intersection of the earth sciences/sustainability and data", "Data-informed food, energy, water and Earth sciences policy discussions" and "Frameworks for helping the scientific and KDD communities to work together".

2.2 Data Mining and AI for Conservation

This workshop focused on how data mining and AI can inform conservation efforts from theory to practice, and from understanding the underlying processes and dynamics to making effective decisions and setting policies. There are increasingly large amounts of data relevant to conservation, including sensor data such as GPS trajectories from collars on animals, camera trap photos, crowdsourced photos and observations, and imagery from UAVs and satellites, DNA

Figure 1: Earth Day workshops and the United Nations goals they relate to.

samples, as well as other geospatial data and data generated by climate models. The organizers brought together researchers in the computational areas of mining images and text, network science, machine learning, planning and optimization, and vision, as well as researchers interested in applications related to climate, environment, ecology and wildlife conservation to discuss use cases of data mining and AI in the nature conservation area.

2.3 Urban Computing

The objective of the Urban Computing workshop was to provide professionals, researchers, and technologists a single forum where they could discuss and share the state-of-the-art of the development and applications related to urban computing, present their ideas and contributions, and set future directions in innovative research for urban computing. At KDD 2019, it targeted people who were interested in sensing/mining/understanding urban data so as to tackle challenges in cities and help better formulate the future of cities. The workshop had two invited talks as well as paper presentations and posters.

3. SESSIONS AND PANEL DISCUSSION

The afternoon plenary sessions were grouped into two themes, each of which had a keynote presentation and panel discussion.

3.1 Importance of Earth Data Sets and Use Cases

This session explored the tremendous value of Earth data for civil society, prosperity and good governance via a keynote and a panel discussion.

Earth data (e.g., remote sensing imagery, GPS time service, location traces) has already transformed our lives by improving the monitoring of global weather and agriculture for early warning of hurricanes and inclement weather as well as food shortage risks due to crop stresses or failures. Further, with two billion [9] receivers in use for location and time services, GPS has become a critical infrastructure of the world economy for use cases ranging from precision agriculture to navigation, ride sharing and smart cities. These success stories are only a beginning and many transformative opportunities lie ahead. For example, the 2011 Mckinsey Big Data report [12] estimated that location data will generate about $600B annually by 2020. In addition, a 2019 U.S. National Academy report projects $1.6T in savings for energy generation and use from Earth data by 2035 [13]. Furthermore, government and industry have recently started major initiatives such as the NASA Earth Exchange [7], Amazon's Earth on AWS [3], Google Earth Engine [4], Microsoft's AI for Earth [6], and NSF's Navigating the New Arctic [8] for meeting the grand challenges facing our planet.

The keynote speaker, Ramakrishna Nemani from NASA Earth Exchange, shared details about the Earth data available at NASA Earth Exchange. He also shared that he was a data mining skeptic for many years and that this was the first time he participated in a data mining meeting. An audience member asked why he had changed his view of the data mining field. Dr. Nemani explained that the recent growth of nano-satellites and availability of satellite imagery data on cloud-computing platforms (Table 2) has overwhelmed the human capacity to analyze the remote sensing data and he sees value in the use of automated methods to help assist human resources. Further, he finds hope in the growing interest in the data mining community to address the issues of interpretability of results. He also expressed enthusiasm for greater involvement of this community in future Earth Days if SIGKDD chooses to continue it.

The keynote address was followed by a panel discussion on four major questions: The first question focused on society and had three parts: 1. What is the societal significance of Earth datasets and what are their most important use cases? What is the annual [economic] value of Earth data expected to be in 2030 or 2040? What is the role of Earth data in good governance? The remaining questions were science-oriented and asked panelists to discuss 2. important types, sources, and accessible repositories of Earth data, 3. unique needs of its use cases, and 4. strengths and weaknesses of current data mining techniques for Earth Data.

Panelist Jennifer Marsman (Microsoft AI for Earth) shared details of a $50M initiative at Microsoft making cloud based resources available for data mining to address the problems facing the Earth. Dennis Pamlin (RISE Research Institutes of Sweden Holding AB) challenged the audience to address vexing sustainability problems where optimizing for one problem (e.g., food production) may make things worse for other problems (e.g. water quality). He urged data miners to look for solutions which help multiple problems at the same time. Rob Bocheneck (Axiom Data Science for NOAA Integrated Ocean Observing System) shared a local perspective describing sensor networks monitoring Alaska seas and processing those using HPC resources.

Fragile Earth Workshop		
1	Snehal More, Anuj Karpatne, et al.	Deep Learning for Forest Plantation Mapping in Godavari Districts of Andhra Pradesh, India
2	Jiri Navratil, Alan King, et al.	Accelerating Physics-Based Simulations Using Neural Network Proxies: An Application in Oil Reservoir Modeling
3	Nishant Yadav, Kate Duffy et al.	Deep Learning Based Quantitative Precipitation Nowcasting
4	Xiaowei Jia, Jared Willard, et al.	Physics Guided Machine Learning for Modeling Engineered and Natural Systems
5	Thomas Uriot	Learning with Sets in Multiple Instance Regression Applied to Remote Sensing
6	Kate Duffy, Thomas Vandal, et al.	DeepEmSat: Deep Emulation for Satellite Data Mining
7	Adrian Albert, Emanuele Strano, et al.	Spatial sensitivity analysis for urban land use prediction with physics-constrained conditional generative adversarial networks
AI for Conservation Workshop		
1	Elizabeth Bondi, Raghav Jain, et al.	Data With a BIRDSAI View: Detection and Tracking in Aerial Thermal Infrared Videos for Conservation
2	Sam McKennoch, Paul Albee, et al.	Towards Automated Metrics for Marine Mammal Health from Aerial Images
3	Sara Beery, Dan Morris, et al.	Efficient Pipeline for Camera Trap Image Review
4	Lily Xu, Shahrzad Gholami, et al.	Stay Ahead of Poachers: Illegal Wildlife Poaching Prediction and Patrol Planning Under Uncertainty with Field Test Evaluations
5	Taoan Huang, Tianyu Gu, et al.	Green Security Game with Community Engagement
6	Guido Muscioni, Riccardo Pressiani, et al.	A Framework for Identifying Group Behavior of Wild Animals
7	Ankush Khandelwal, Anuj Karpatne, et al.	GLADD-R: A new Global Lake Dynamics Database for Reservoirs created using Machine Learning and Satellite Data
Urban Computing Workshop		
1	Shanaka Perera, Theo Damoulas, et al.	Modelling Business Rates in England with Big Spatial Data
2	Phillip Taylor, Nathan Griffiths, et al.	Data mining and compression: where to apply it and what are the effects?
3	Ariel Noyman, Ronan Doorley, et al.	What's your MoCho? Real-time Mode Choice Prediction Using Discrete Choice Models and a HCI Platform
4	Sirui Song, Tong Xia, et al.	UrbanRhythm: Revealing Daily Urban Dynamics Hidden in Mobility Data
5	Zhuojie Huang	Vision Paper: From Data Science to Blockchain - Analytics in Cross-Border Logistics
6	Deeksha Goyal, Albert Yuen, et al.	Traffic Control Elements Inference using Telemetry Data andConvolutional Neural Networks
7	Arpan Man Sainju, Zhe Jiang	Mapping Road Safety Features from Streetview Imagery: A DeepLearning Approach

Table 1: Posters exhibited as part of the Earth Day symposium.

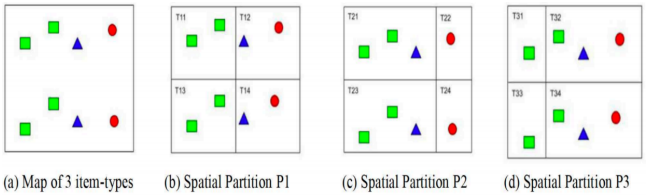
Datasets	Google Earth Engine	NEX	Earth on AWS
Elevation, Landsat, LOCA, MODIS, NAIP	X	X	X
NOAA	X		X
AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1	X	X	
IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)			X
CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI, WHRC, WorldClim, WorldPop, WWF	X		
BCCA, FLUXNET		X	

Table 2: Earth datasets available on the cloud.

3.2 Earth Day Related Data Mining challenges and Opportunities

This session focused on the special challenges and opportunities that Earth data poses for data science.

Data mining methods have found success in analyzing many complicated systems, such as e-commerce, and use cases explored in the Earth Day aligned SIGKDD workshops. However, many questions remain open due to the unique Earth data challenges such as spatio-temporal auto-correlation, heterogeneity, scale-dependence, measurement errors, modifiable areal unit problem, etc [11, 1, 15]. A recent paper in Geo-Physical Letters [10] noted that "failure to account for dependence between [Physical] models, variables, locations and seasons yields misleading results". Additional challenges have been noted in recent community papers from the NSF IS-GEO Research Coordination Network [11] and University Consortium for Geographic Information Science [1]. Court debates on gerrymandering [5] also raise transparency concerns for the risk of altering statistical results by changing the choice of spatial partitions.



(a) Map of 3 item-types (b) Spatial Partition P1 (c) Spatial Partition P2 (d) Spatial Partition P3

Spatial Partitioning Definition	P1	P2	P3
Transactions	T11, T12, T13, T14	T21, T22, T23, T24	T31, T32, T33, T44
Associations with support >= 0.5	(▲ ●)	(■ ▲)	(■ ▲ ●)

Figure 2: Gerrymandering and association rules. Consider the spatial Data in Figure (a) along with 3 alternative partitions in Figures (b), (c) and (d).

Prof. Harvey Miller shared highlights from a recent (July 2019) NSF workshop on mobility and cities exploring urban mobility observatories along with data mining challenges such as gerrymandering (i.e. modifiable areal unit problem) (Figure 2). In other words, results of classical data mining methods (e.g., association rules) can be gamed by changing the boundaries of spatial partitions defining transactions.

The following questions were asked to the panel: 1. List knowledge gaps between Earth data mining needs and data mining state of the art. 2. What new research is needed to fill the knowledge gaps? 3.What are the grand challenges for data mining with respect to analyzing Earth data? 4. How may data mining methods address the modifiable areal unit problem? Spatial bias in data? Social feedback loops increasing spatial bias? 5. Is there a need for SIGKDD community action? If so, suggest community actions.

Panelist Prof. Dan Griffith (UT-Dallas) described spatial data science challenges such as spatial auto-correlation.

Prof. Anuj Karpatne (Virginia Tech) shared his recent work on physics guided data mining to address challenges faced when mining Earth data. Prof. Tanya Berger-Wolf (University of Illinois at Chicago) shared her perspective on leveraging social media data to estimate populations of endangered species. Audience questions included how to address spatial bias in data. For example, social media reports of animals are not covering large parts of Alaska where few humans visit. Panelists brainstormed on alternative ways to address spatial bias including comparing social media reports with gold standard data.

The discussions begun at this session continued beyond the Earth Day event and spanned the whole SIGKDD conference. Email exchanges indicate an increased awareness and interest from the data science community in contributing to Earth-related research.

4. CONCLUSIONS

Earth Day at KDD 2019, the first event of its kind, dedicated a full day to the use of data science for Earth research. The event started discussions among stakeholders and participants who are looking forward to growing the Earth sciences community. Follow-up activities, i.e. "All Hands Meeting of Midwest Big Data Hub" in Oct. 2019 (A Panel titled "Role of Big Data in the American AI Initiative" and a Community Whitepaper to inform science policy) and "American Association for Advancement of Sciences" meeting in Feb. 2020 (Session titled "Using Computing to Sustainably Feed a Growing Population"), were organized to continue the discussions among stakeholders.

5. ACKNOWLEDGMENTS

We would like to thank the Earth Day as well as the workshop co-organizers for their valuable contributions. Also, we would like to thank KDD 2019 co-chairs for their support on organizing the Earth Day event as a full day of Earth-related synergistic activities.

6. ADDITIONAL AUTHORS

Hui Xiong (Rutgers, hxiong@rutgers.edu), Jieping Ye (U. of Michigan, jpye@med.umich.edu), Xun Zhou (U. of Iowa, xun-zhou@uiowa.edu), Ramasamy Uthurusamy (GM, samy @gm.com), Chid Apte (IBM Research, apte@us.ibm.com), Naoki Abe (IBM Research, nabe@us.ibm.com), Vani Mandava (Microsoft, vanim@ microsoft.com), Meredith Lee (UC Berkeley, mmlee@berkeley. edu), Lea Shanley (Wilson Center, lea.shanley@ wilsoncenter. org) Vipin Kumar (U. of Minnesota, kumar001@umn.edu) and Yiqun Xie (U. of Minnesota, xiexx347@umn.edu).

7. REFERENCES

[1] Bringing the geospatial perspective to data science degrees and curricula. https://www.ucgis.org/assets/docs/UCGIS-Statement-on-Data-Science-Summer2018.pdf. Accessed: 2019-09-03.

[2] Earth day at kdd 2019. https://www.kdd.org/kdd2019/special-days/earth-day. Accessed: 2019-09-03.

[3] Earth on amazon aws. https://aws.amazon.com/earth/. Accessed: 2019-09-03.

[4] Google earth engine. https://earthengine.google.com/. Accessed: 2019-09-03.

[5] Justices display divisions in new cases on voting maps warped by politics. https://www.nytimes.com/2019/03/26/us/politics/gerrymandering-supreme-court.html. Accessed: 2019-09-03.

[6] Microsoft ai for earth. https://www.microsoft.com/en-us/ai/ai-for-earth. Accessed: 2019-09-03.

[7] Nasa earth exchange. https://nex.nasa.gov/nex/. Accessed: 2019-09-03.

[8] Navigating the new arctic. https://www.nsf.gov/news/special_reports/big_ideas/arctic.jsp. Accessed: 2019-09-03.

[9] The world economy runs on gps. it needs a backup plan. https://www.bloomberg.com/news/features/2018-07-25/the-world-economy-runs-on-gps-it-needs-a-backup-plan. Accessed: 2019-09-03.

[10] P. M. Caldwell, C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, 41(5):1803–1808, 2014.

[11] Y. Gil, S. A. Pierce, H. Babaie, A. Banerjee, K. Borne, G. Bust, M. Cheatham, I. Ebert-Uphoff, C. Gomes, M. Hill, et al. Intelligent systems for geosciences: an essential research agenda. *Communications of the ACM*, 62(1):76–84, 2018.

[12] J. Manyika. Big data: The next frontier for innovation, competition, and productivity. *http://www. mckinsey. com/Insights/MGI/Research/Technology and Innovation/ Big data The next frontier for innovation*, 2011.

[13] N. A. of Sciences Space Studies Board. *Thriving on our changing planet: A decadal strategy for Earth observation from space*. National Academies Press, 2019.

[14] U. SDGs. United nations sustainable development goals. *UN. Org*, 2015.

[15] Y. Xie, E. Eftelioglu, R. Ali, X. Tang, Y. Li, R. Doshi, and S. Shekhar. Transdisciplinary foundations of geospatial data science. *ISPRS International Journal of Geo-Information*, 6(12):395, 2017.