Location Powers: Data Science Summit

November 13 and 14, at Google, Mountain View, CA
Concept Note: 2019-05-06

The explosive availability of data about nearly every aspect of human activity along with revolutionary advances in computing technologies is transforming geospatial data science. The shift from data-scarce to data-rich environment comes from mobile devices, remote sensing and the Internet of Things. Nearly all of this data has components of location and time. Innovations in cloud computing and big data provides methods to perform data analytics at exceedingly large scale and speed. The development of intelligent systems using knowledge models and their impact on our insights and understanding will be focus of the Location Powers: Data Science Summit.

The Location Powers Summit series brings together industry, research and government experts from across the globe into an interactive discussion that assess the current situation and produces recommendations for future technology innovations and standards development. The Location Powers Summits are key to the technology innovation promoted by the Open Geospatial Consortium (OGC).

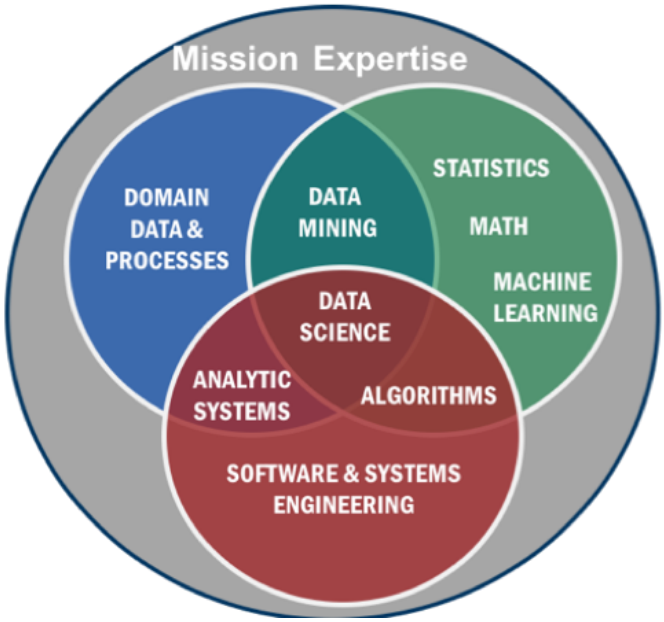
The Location Powers: Data Science Summit will convene experts on data science, machine learning, artificial intelligence, cloud computing, remote sensing and GIS to provide a technology basis. Participation by leaders in social sciences, business development and government policy will lead to recommendations that have meaningful outcomes from the geospatial data science developments.

The remainder of this note provides an extended definition of Geospatial Data Science. An OGC Tech Note will be developed based on the Summit. The Geospatial Data Science Tech Note will capture the content of the Summit and provide a basis for further action in OGC and beyond.

Geospatial Data Science – An extended definition

- Data Science and Big Data

Data Science in the context of Big Data systems includes: 1) mathematical and computer science foundations in statistics and machine learning; along with 2) software and systems engineering methods to handle large data volumes and innovative query and analytics techniques; and, in some extended definitions, may include 3) domain data and processes.



Data Science - from NIST Big Data interoperability Framework, Volume 1 – Definitions

The emergence of Data Science concepts and motivation can be traced to Jim Grey's concepts captured in The Fourth Paradigm: Data-Intensive Scientific Discovery, by Tony Hey, Stewart Tansley, and Kristin Tolle. This book surveys opportunities and challenges for data-intensive science to prepare for the data deluge of a "sensors everywhere" data infrastructure supporting a fourth paradigm of scientific research based on "Data Exploration". Alternatively, the methods of data exploration can be seen as emerging from statistics and a fifty-year history of Data Science [David Donoho]. Success has many parents.

More recently, we see the emergence of a new field — Data Science — that focuses on the processes and systems that enable us to extract knowledge or insight from data in various forms and translate it into action. In practice, data science has evolved as an interdisciplinary

field that integrates approaches from such data-analysis fields as statistics, data mining, and predictive analytics and incorporates advances in scalable computing and data management. [Realizing the Potential of Data Science - ACM Berman 2018

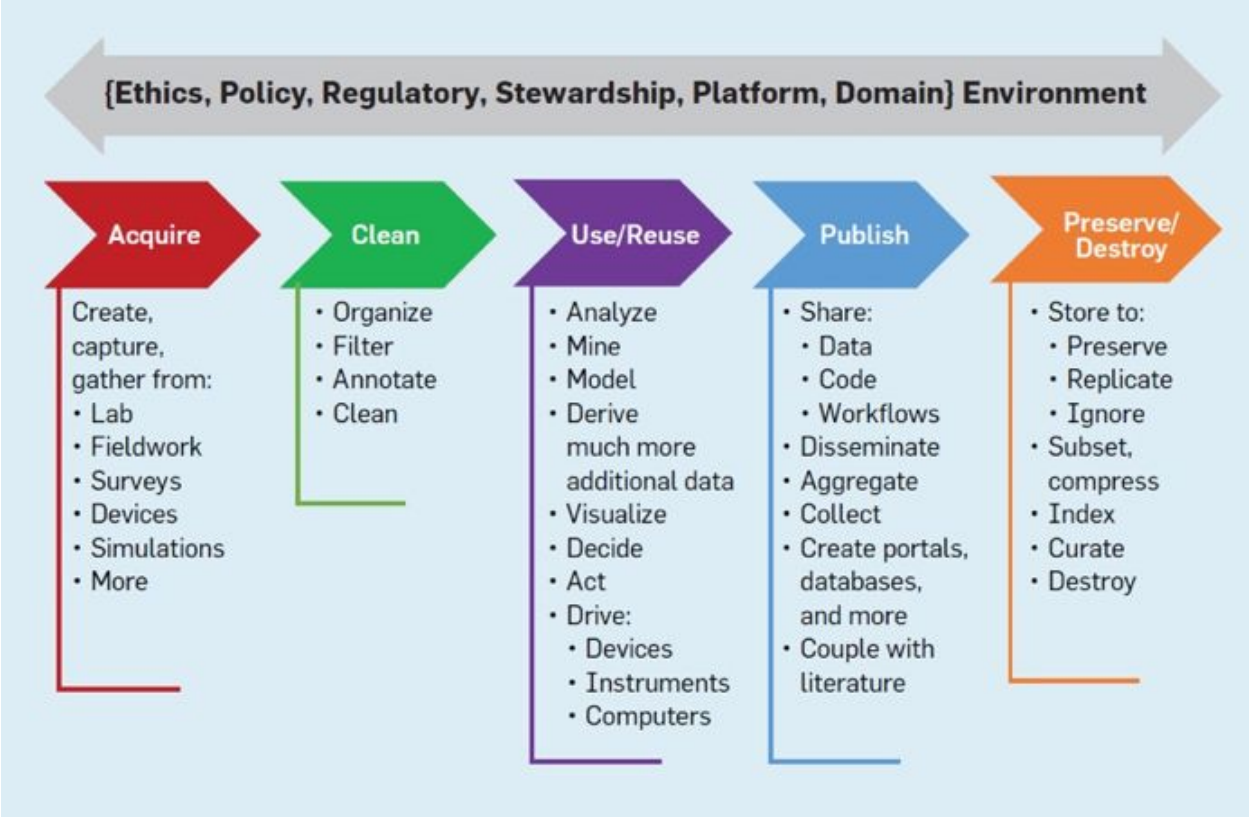


Figure. Data life cycle from the *Realizing the Potential of Data Science Report.*[2]

Methods in Data Science and defined in more limited discussions focuses on the mathematical and computer science algorithm-based techniques [Foundations of Data Science]

- High-Dimensional Space
- Best-Fit Subspaces and Singular Value Decomposition (SVD)
- Random Walks and Markov Chains
- Machine Learning
- Algorithms for Massive Data Problems
- Clustering
- Random Graphs
- Topic Models, Hidden Markov Models, and Graphical Models

Extending data science beyond math and computer science brings in Theory-Guided Data Science. [Karpatne, et.al.] that aims to leverage the wealth of scientific knowledge for improving the effectiveness of data science models in enabling scientific discovery.

Responding to these trends, Universities have established Data Science Programs. For example, the University of Michigan's Data Science Initiative was established to focus on: This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications."

- · Geospatial Data Science

Applying Data Science to Geospatial Information is producing tremendous results. Geospatial information is both similar and different than other types of information considered in Data Science. Some of the results are geospatial data science are by extending existing data science methods to geospatial information. Some results are by defining new methods for examining geospatial information that leads to extensions in data science.
Geospatial information is experiencing the data explosion of mobile devices, remote sensing and the Internet of Things perhaps more than other fields as all of these data types include location, spatial and temporal information.

During SciDataCon2016, OGC organized the session "Geospatial Data and Key Characteristics of Geospatial Data Analysis and Science: The way forward." The session targeted geo-spatial data as a fundamental base layer for data science and analysis. It addressed geo-spatial properties of data and discusses which new models need to be developed to enhance the level of interoperability from syntactical to semantic interoperability. The goal was to compare various approaches, to discuss the advantages and disadvantages and their relevance, capabilities, and efficiencies in the context of data analysis.

The ESIP Federation has defined "Earth Science Data Analytics as: Process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data encompassing a variety of data types to uncover patterns, correlations and other information, to better understand our Earth. [ESIP 2016]

In 2018, these topics were addressed in the CGA Conference: Illuminating Space and Time in Data Science. The emergence of Data Science has provided a renewed opportunity to consider the importance of spatial relationships at the heart of these smart sensors and Internet of Things (IoT). Indeed, space and time are core properties of 'big data', so called, and spatiotemporal analysis is inherently an important facet in Data Science. From satellite images to social media streams, from census and parcels to records of trade, food, energy, climate, disease, crime, conflicts, etc., big data with space and time signatures are essential for understanding our world and responding to its challenges. The conference aimed at bringing together mainstream data scientists and geographic information scientists, to review the status of both fields, explore commonalities between the two, and identify the relevance of space and time in Data Science.

Geospatial Data Science Techniques include the following:

- Data Representation: Maps, Features, Coverages
- Geographically-distributed; Cloud, Fog, Edge Computing
- Sensor Observations and Feature Extraction
- Geospatial fusion and conflation
- Data Mining: i.e., hotspot detection, colocation detection, prediction, outlier detection and teleconnection detection
- Machine learning, in particular on geospatial imagery
- Geo-semantics and linked data

Data Lifecycle for Geospatial Data Science (update Figure from OGC Big Data White Paper).

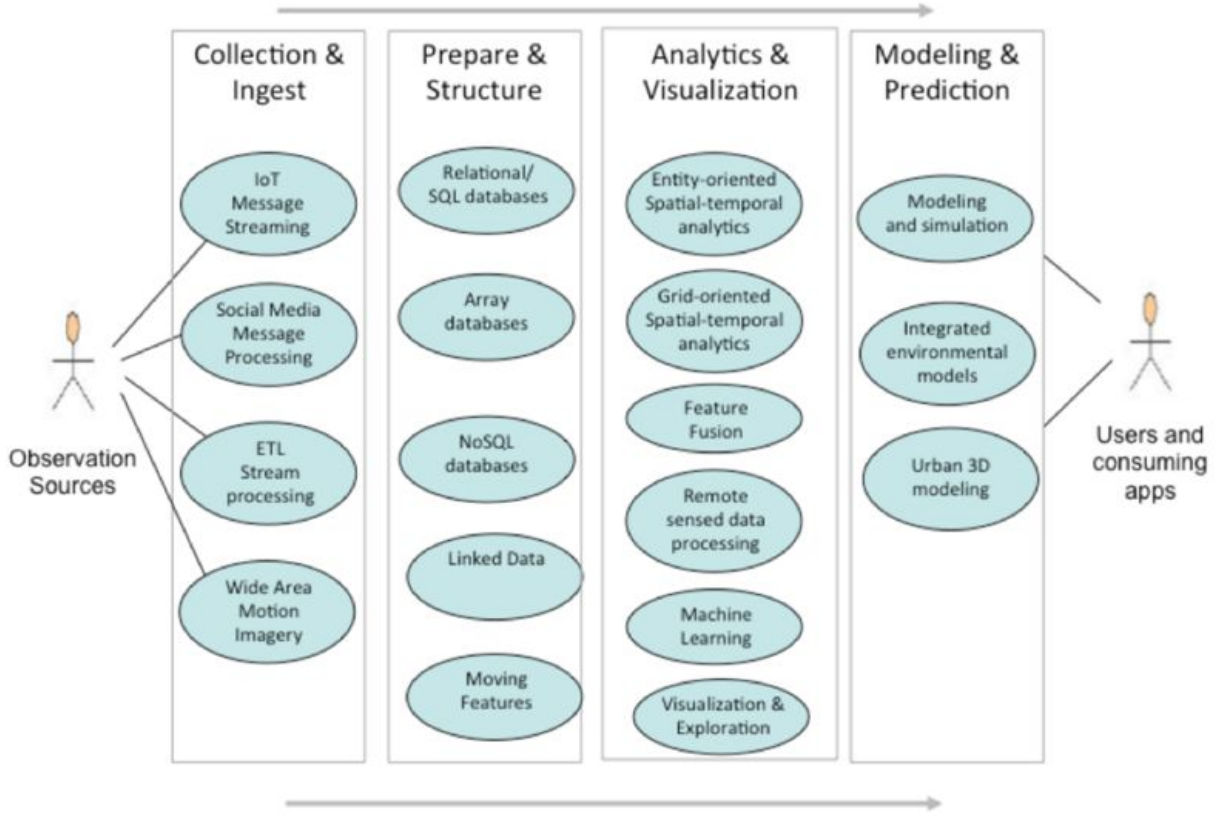


Figure. Geospatial Data Science Workflow

- Application and Social Impact of Geospatial Data Science

So what good is Geospatial Data Science? Develop a list of applications, such as:
- Smart Cities
- Transportation
- Geosciences
- more

Social impact
- ACM Statement on Algorithmic Transparency and Accountability
- New York City Algorithm Task Force
- Reproducibility

References

NIST Big Data interoperability Framework (NBDIF), Volume 1: Definitions, https://bigdatawg.nist.gov/V3_output_draft_docs.php

Longbing Cao. 2017. Data science: challenges and directions. **Communications of the ACM**, 60, 8 (July 2017), 59-68. DOI: https://doi.org/10.1145/3015456

Hey, Tony and Tansley, Stewart and Tolle, Kristin, The Fourth Paradigm: Data-Intensive Scientific Discovery, Published by Microsoft Research, October 2009 ISBN: 978-0-9825442-0-4

Realizing the Potential of Data Science - ACM Berman 2018. Https://dl.acm.org/citation.cfm?Id=3188721

REALIZING THE POTENTIAL OF DATA SCIENCE - NSF. Https://www.nsf.gov/cise/ac-data-science-report/ciseacdatasciencereport1.19.17.pdf

50 years of Data Science, David Donoho, 2015

Foundations of Data Science, Avrim Blum, John Hopcroft, and Ravindran Kannan, January 2018

Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. Karpatne, et.al. https://arxiv.org/abs/1612.08544

“Geospatial Data and Key Characteristics of Geospatial Data Analysis and Science: The way forward, SciDataCon2016, Session Organisers: Luis Bermudez , Ingo Simonis. https://www.scidatacon.org/2016/sessions/99/

Wang, Shaowen “A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis,” Annals of the Association of American Geographers, 2010/06/25, doi: 10.1080/00045601003791243

“Transdisciplinary Foundations of Geospatial Data Science.” Xie, et.al., University of Minnesota December 5, 2017

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (May 2017), 84-90. DOI: https://doi.org/10.1145/3065386

D. Lunga, et.al., "Domain-Adapted Convolutional Networks for Satellite Image Classification: A Large-Scale Interactive Learning Workflow," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 3, pp. 962-977, March 2018.
doi: 10.1109/JSTARS.2018.2795753

“Deep Learning for Classification Tasks on Geospatial Vector Polygons,” Rein van 't Veer, Peter Bloem, Erwin Folmer. https://arxiv.org/abs/1806.03857

Yolanda Gil, et.al. 2018. Intelligent systems for geosciences: an essential research agenda. *Commun. ACM* 62, 1 (December 2018), 76-84. DOI: https://doi.org/10.1145/3192335

L. Wang, B. Guo and Q. Yang, "Smart City Development With Urban Transfer Learning" in Computer, vol. 51, no. 12, pp. 32-41, 2018. doi: 10.1109/MC.2018.2880015