Open Geospatial Consortium

Submission Date: <2020-05-20>

Approval Date: <yyyy-mm-dd>

Publication Date: <yyyy-mm-dd>

External identifier of this OGC® document: http://www.opengis.net/doc/{doc-type}/{standard}/{m.n}

Internal reference number of this OGC® document: 20-001r1

Category: OGC® White Paper

Editor: George Percivall

Geospatial Data Science

Document type:	OGC® White Paper
Document subtype:	
Document stage:	Draft
Document language:	English

License Agreement

Permission is hereby granted by the Open Geospatial Consortium, ("Licensor"), free of charge and subject to the terms set forth below, to any person obtaining a copy of this Intellectual Property and any associated documentation, to deal in the Intellectual Property without restriction (except as set forth below), including without limitation the rights to implement, use, copy, modify, merge, publish, distribute, and/or sublicense copies of the Intellectual Property, and to permit persons to whom the Intellectual Property is furnished to do so, provided that all copyright notices on the intellectual property are retained intact and that each person to whom the Intellectual Property is furnished agrees to the terms of this Agreement.

If you modify the Intellectual Property, all copies of the modified Intellectual Property must include, in addition to the above copyright notice, a notice that the Intellectual Property includes modifications that have not been approved or adopted by LICENSOR.

THIS LICENSE IS A COPYRIGHT LICENSE ONLY, AND DOES NOT CONVEY ANY RIGHTS UNDER ANY PATENTS THAT MAY BE IN FORCE ANYWHERE IN THE WORLD.

THE INTELLECTUAL PROPERTY IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE DO NOT WARRANT THAT THE FUNCTIONS CONTAINED IN THE INTELLECTUAL PROPERTY WILL MEET YOUR REQUIREMENTS OR THAT THE OPERATION OF THE INTELLECTUAL PROPERTY WILL BE UNINTERRUPTED OR ERROR FREE. ANY USE OF THE INTELLECTUAL PROPERTY SHALL BE MADE ENTIRELY AT THE USER'S OWN RISK. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR ANY CONTRIBUTOR OF INTELLECTUAL PROPERTY RIGHTS TO THE INTELLECTUAL PROPERTY BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM ANY ALLEGED INFRINGEMENT OR ANY LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR UNDER ANY OTHER LEGAL THEORY, ARISING OUT OF OR IN CONNECTION WITH THE IMPLEMENTATION, USE, COMMERCIALIZATION OR PERFORMANCE OF THIS INTELLECTUAL PROPERTY.

This license is effective until terminated. You may terminate it at any time by destroying the Intellectual Property together with all copies in any form. The license will also terminate if you fail to comply with any term or condition of this Agreement. Except as provided in the following sentence, no such termination of this license shall require the termination of any third party end-user sublicense to the Intellectual Property which is in force as of the date of notice of such termination. In addition, should the Intellectual Property, or the operation of the Intellectual Property, infringe, or in LICENSOR's sole opinion be likely to infringe, any patent, copyright, trademark or other right of a third party, you agree that LICENSOR, in its sole discretion, may terminate this license without any compensation or liability to you, your licensees or any other party. You agree upon termination of any kind to destroy or cause to be destroyed the Intellectual Property together with all copies in any form, whether held by you or by any third party.

Except as contained in this notice, the name of LICENSOR or of any other holder of a copyright in all or part of the Intellectual Property shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Intellectual Property without prior written authorization of LICENSOR or such copyright holder. LICENSOR is and shall at all times be the sole entity that may authorize you or any third party to use certification marks, trademarks or other special designations to indicate compliance with any LICENSOR standards or specifications. This Agreement is governed by the laws of the Commonwealth of Massachusetts. The application to this Agreement of the United Nations Convention on Contracts for the International Sale of Goods is hereby expressly excluded. In the event any provision of this Agreement shall be deemed unenforceable, void or invalid, such provision shall be modified so as to make it valid and enforceable, and as so modified the entire Agreement shall remain in full force and effect. No decision, action or inaction by LICENSOR shall be construed to be a waiver of any rights or remedies available to it.

Table of Contents

i. Abstract

This OGC White Paper describes Geospatial Data Science based on the Location Powers: Data Science Summit of November 2019. The white paper provides a description of the presentations and discussions of the summit along with recommendations for OGC activities to advance the field of Geospatial Data Science.

ii. Keywords

The following are keywords to be used by search engines and document catalogues.

ogcdoc, OGC document, Data Science, Analytics, Statistics, Artificial Intelligence, Machine Learning, Edge Computing, Knowledge-based Models, Data Management, IT Ethics, Heterogenous computing

iii. Preface

Geospatial Data Science is defined in this white paper as “The art and craft of people leveraging technology to create value out of data using location and time.” The components of geospatial data science are data, tools, applications, ethics, and emerging trends. The data component is composed of discussions about big geospatial data; data scientists, teams and process; and data management. The tools component is composed of discussions about geospatial representations and analytics, the application of machine learning to geospatial, and knowledge-based models to support decision making.

An objective of the white paper to serve as a basis for the promotion of geospatial data science within and external to OGC. OGC has a role to conduct activities that will advance innovation and standardization in geospatial data science. The overall objective is to enable beneficial use of geospatial information in humanities critical decisions.

iv. Submitting organizations

This document is based on material from organizations that participated in the Location Powers: Data Science Summit as listed in Annex A.

v. Submitters

All questions regarding this submission should be directed to the editor: George Percivall, Open Geospatial Consortium

Chapter 1. Overview of White Paper

Geospatial Data Science has been identified as an important technology development trend by the Open Geospatial Consortium (OGC). The OGC Technology Forecasting activity began focusing on data science as an outcome of the development of the Big Geospatial Data topic area. Both Big Data and Data Science have been topics in recent Location Powers Summits.

The Location Powers: Data Science Summit (LP_DS) organized by OGC was held on November 13 and 14, 2020, hosted by Google in Mountain View, CA. This Geospatial Data Science White Paper captures the content of the Summit and provide a basis for further action in OGC and beyond.

Location Powers Summit brings together industry, research and government experts from across the globe into an interactive discussion that assesses 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 OGC.

The Location Powers: Data Science Summit convened experts on data science, machine learning, artificial intelligence, cloud computing, remote sensing and GIS to assess the current situation of geospatial data science. Participation by leaders in social sciences, business development, government policy and information technology led to recommendations with meaningful outcomes for geospatial data science development.

The LP_DS Summit considered the explosive availability of data about nearly every aspect of human activity along with revolutionary advances in computing technologies that 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 was the focus of the LP_DS.

A summary of the topics discussed in the LP_DS is shown in the figure:

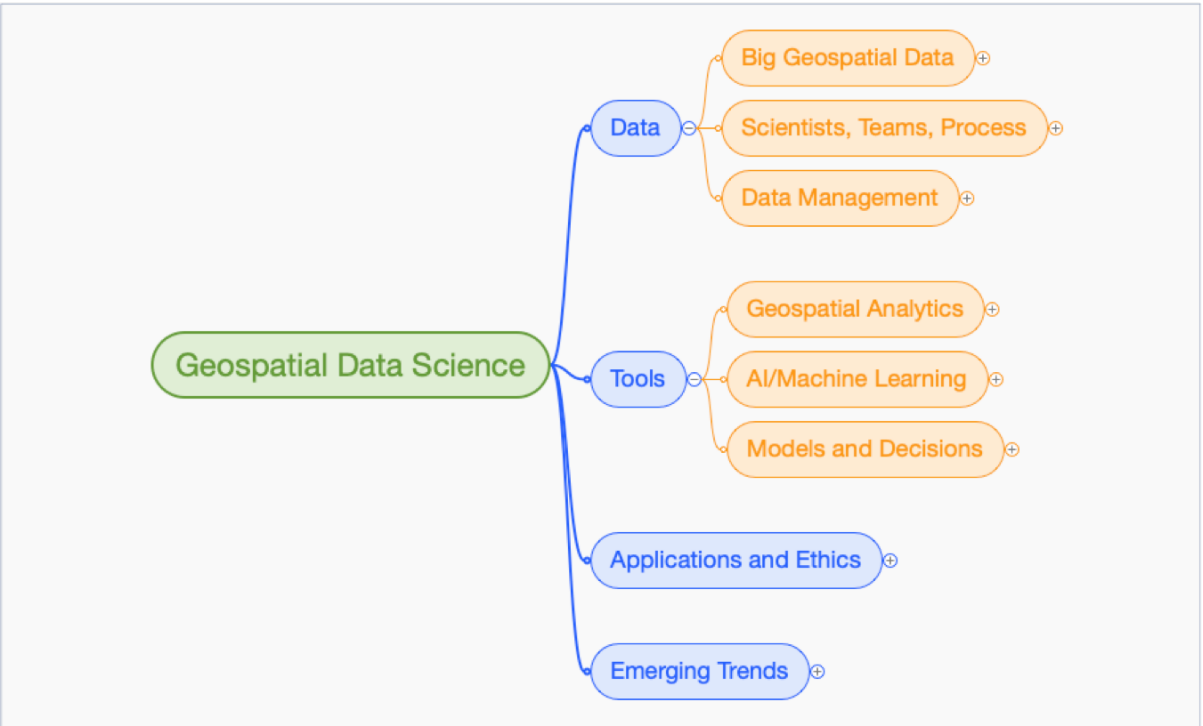


Figure 1. Geospatial Data Science

This White Paper is organized as follows:

- Data Topics
  - Big Geospatial Data (Clause 3)
  - Data Scientists, Teams, Process (Clause 4)
  - Data Management (Clause 5)
- Tools
  - Geospatial Representations and Analytics (Clause 6)
  - AI and Machine Learning (Clause 7)
  - Models and Decisions (Clause 8)
- Data Science Applications and Ethics (Clause 9)
- Emerging Trends (Clause 10)

The Emerging Trends are: Edge Computing and Heterogeneous Computing

An Annex provides information about the summit including: the agenda and the organizations that participated in the Summit.

Chapter 2. Overview of Geospatial Data Science

This definition was developed and repeated in several presentations and discussion sessions of the Location Powers Data Science Summit (LP_DS):

```
Geospatial Data Science is "The art and craft of people leveraging technology
to create value out of data using location and time."
```

To set the context for LP_DS, a definition for Data Science in the context of Big Data systems coming from NIST was considered. The NIST Big Data Interoperability Framework defines Data Science as the extraction of useful knowledge directly from data through a process of discovery, or of hypothesis formulation and hypothesis testing. The NIST document goes on to identify Data Science Sub-disciplines as 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.

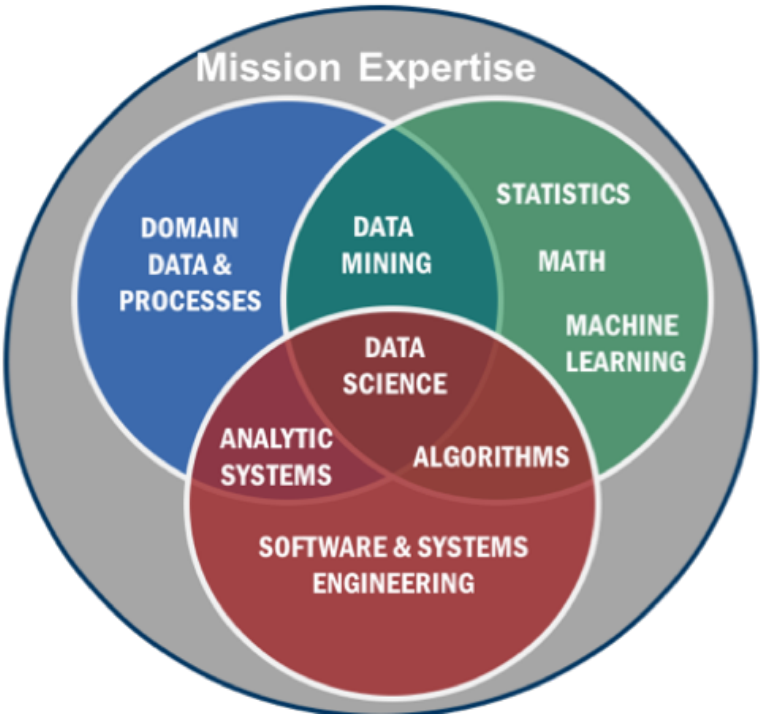


Figure 2. Data Science from NIST Big Data interoperability Framework

Applying Data Science in the context of Geospatial Information is producing tremendous results. 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.

The Location Powers: Data Science Summit expanded beyond the topics listed above leading to this outline of key topics in Geospatial Data Science: Data, Tools, Applications, and Trends.

- Data: It is obvious but important to state that Data is a core topic of data science. The availability

of increasing availability of data triggered new possible analysis. Geospatial Data which has always been big data provides opportunities for analytics in data science. Therefore the opening discussion of data is about Big Geospatial Data (Clause 3). For data science to be effective, data scientists needs to work in multi-disciplinary teams with an agile process. These topics are addressed in Data Scientists, Teams, Process (Clause 4). Managing big data requires addressing data policy along with the ecosystems and platforms to manage the data. Cloud-Native data management is providing nimble and novel methods to work with big data. These topics are addressed in Data Management (Clause 5)

- Tools: Working with Big Data requires appropriate tools. As geospatial has always been big data, many of the geospatial analysis methods were data science before the term was introduced. Methods long familiar to the geospatial community along with extensions to those methods are addressed in the clause on Representation and analytics (Clause 6). The third wave of Artificial Intelligence has been lead by machine learning based such as convolutional neural networks. The application of machine learning to big geo data in particular imagery is addressed in AI and Machine Learning (Clause 7). Knowledge based data science depends upon models that are predictive of some portion of the geospatial world. Spatial decision support is supported by knowledge based models. These topics are address in the last tools clause on Models and Decisions (Clause 8).
- Applications and Ethics. Applying Data Science to geospatial data is producing results which were discussed in the summit The Summit discussed nearly a dozen application areas. The applications discussion surfaced need for consideration of ethics regarding Data, Algorithms. (Clause 9)
- Trends that look to be further advancing geospatial data science include Computing at the Edge and Heterogenous Computing. Each of these are addressed in Emerging Trends (Clause 10).

Chapter 3. Data: Big Geospatial Data

The emergence of Data Science concepts and motivation can be traced to Jim Grey's concepts 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". A recurring theme in Location Powers: Data Science summit was that of "telling stories with data". Using stories to explore and understand the data from a domain results in insights not previously available. Data Science can be described as the exploration of big data about a domain.

This Clause addresses topics related to big data for data science.

- Big Data with Location
- Big Data Software Stack
- Big Geo Data Use Cases
- Recommendations

3.1. Big Data with Location

Geospatial data has always been big data was a theme of two Location Powers: Big Data summits and the resulting Big Geospatial Data – an OGC White Paper. The Big Geo Data white paper had these main themes:

- Geospatial data is increasing in volume and variety;
- New Big Data computing techniques are being applied to geospatial data;
- Geospatial Big Data techniques benefit many applications; and
- Open standards are needed for interoperability, efficiency, innovation and cost effectiveness.

The growth of geospatial highlighted in the Big Geo Data White Paper continues and is increasing. Patrick Griffiths, ESA, highlighted this trend during LP_DS. The ESA archives alone will be over 100 Petabytes by 2026.

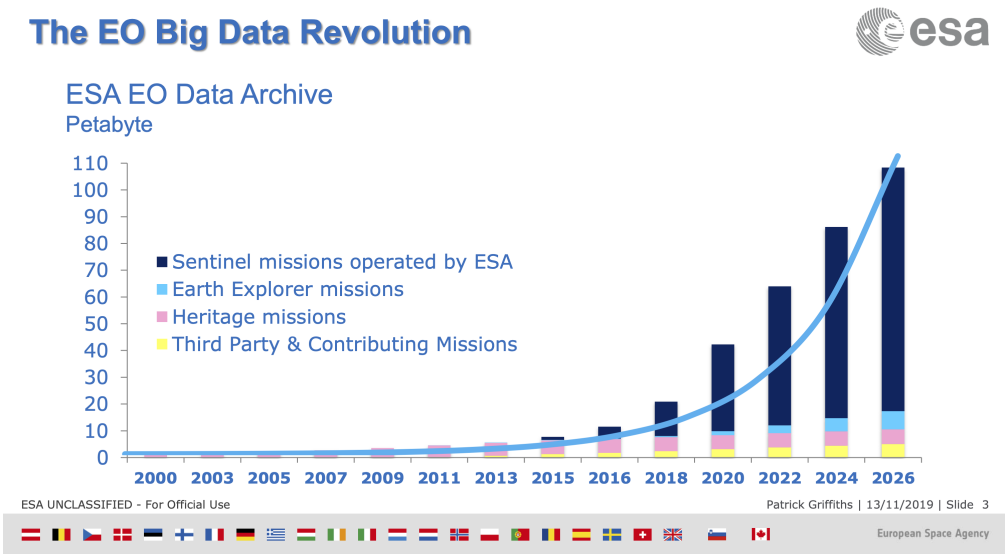


Figure 3. The EO Big Data Revolution

Marc Armstrong, Univ of Iowa, at LP_DS described future satellite constellations that are being planned by different companies including Amazon and SpaceX. SpaceX is planning to deploy 12,000 satellites for communications, military, and scientific purposes. The revisit rate for viewing locations will increase dramatically. BlackSky is proposing 40 to 70 times each day. In addition to the static imagery there's a lot of streaming video that's going to be provided as well.

The Big Geo Data revolution is not only driven by remote sensing from satellites. Philippe Cases, Topio Networks, provided estimates to LP_DS on the magnitude of the data deluge coming from edge devices. All of this Edge Data has components of location and time that can be exploited in data science.

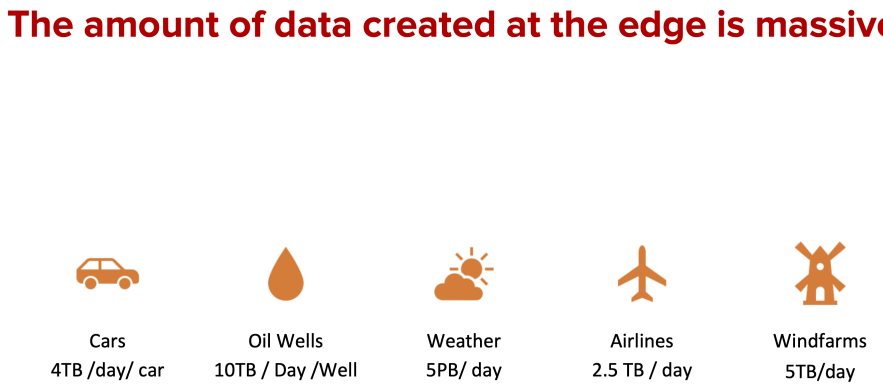


Figure 4. Big Data Revolution from the Edge

It is important to emphasis that this growing data has components of location and time. During LP_DS, Ed Parsons, Google, emphasized the ubiquity of location by introducing the definition of "ambient location."

```
Ambient Location
adjective
denoting or relating to a knowledge of a location that is continuously accessible.
"A smartphone provides the user with an ambient location service."
```

3.2. Big Data Software Stack

The growth of a Big Data Stack drove development of a fundamentally different software computing platform. The birth of the Big Data Stack in late 1990s and early 2000s provided extreme flexibility and scalability in distributed batch applications for data at ever increasing volumes. The Modern Data Architecture provides a good summary of these developments and includes this figure.

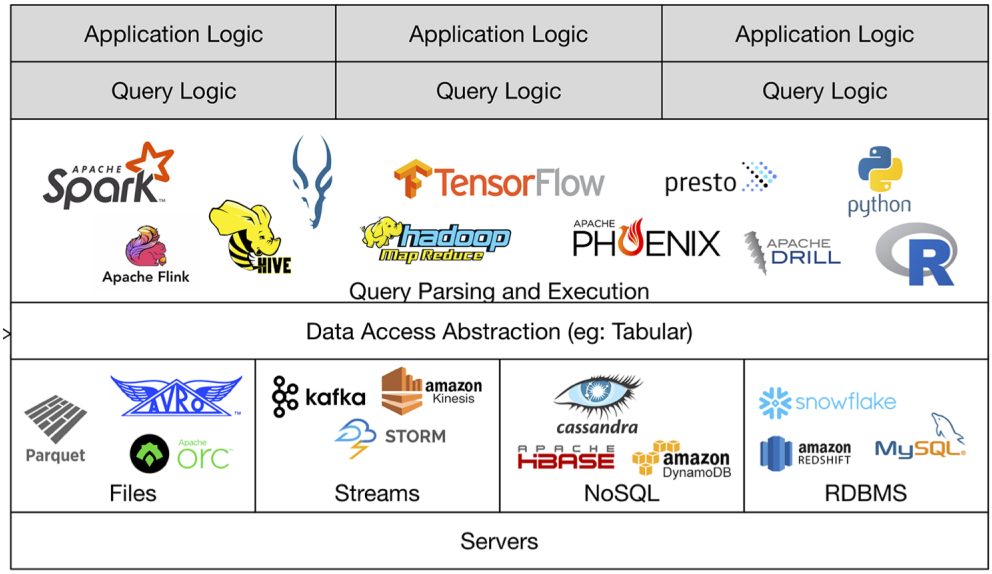


Figure 5. The Birth of the Big Data Stack

At the core of the big data stack was Apache Hadoop, which started in 2006 as a spin-off from Apache Nutch, a web crawler that stemmed from Apache Lucene, the famous open source search engine. The inspiration for this project came from the Google File System and a distributed processing framework called MapReduce. These two components combined the extreme flexibility and scalability necessary to develop distributed batch applications in a simple way.

The use of Big Data Stack software for geospatial applications has been the theme of the Geospatial Track at the annual Apache Conference. The Apache Software Foundation has been a focal point for development of packages of the big data stack. These big data software packages have been extended with geospatial functionality and presented in the ApacheCon geospatial track. These items were presented in the ApacheCon 2019 Geospatial Track: GeoSpark built on Apache Spark, Apache Science Data Analytics Platform, GeoMesa on top of Accumulo, HBase, Cassandra, Geospatial Indexing and Search at Scale with Apache Lucene, Realtime Geospatial Analytics with GPUs, RAPIDS, and Apache Arrow

In later clauses of this white paper we will see how the Big Data Stack is important to data management (Clause 5), geospatial analytics (Clause 6), and Machine Learning (Clause 7).

3.3. Big Geo Data Use Cases

Milind Naphade, NVIDIA Metropolis, picked up on the LP_DS theme of big geo data discussing spatial intelligence. Exploiting this growth in data will require both cloud computing and Computing at the Edge (See Clause 10 for more on this emerging trend). Both the volume and the rate at which these data is coming requires pushing the processing closer to source at the edge. This will impact many vertical applications in terms of getting situational awareness.

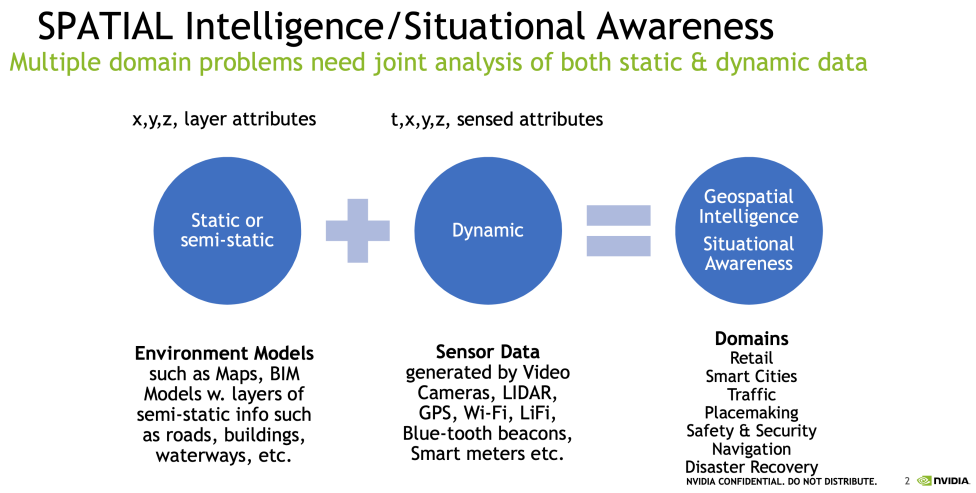


Figure 6. Situational Awareness based on Location

The Big Geospatial Data – an OGC White Paper presented a set of use cases that apply across the application domains. The Use Cases were organized into four groups as shown in the figure. The use cases to the right of the figure provide a motivation for Geospatial Data Science.

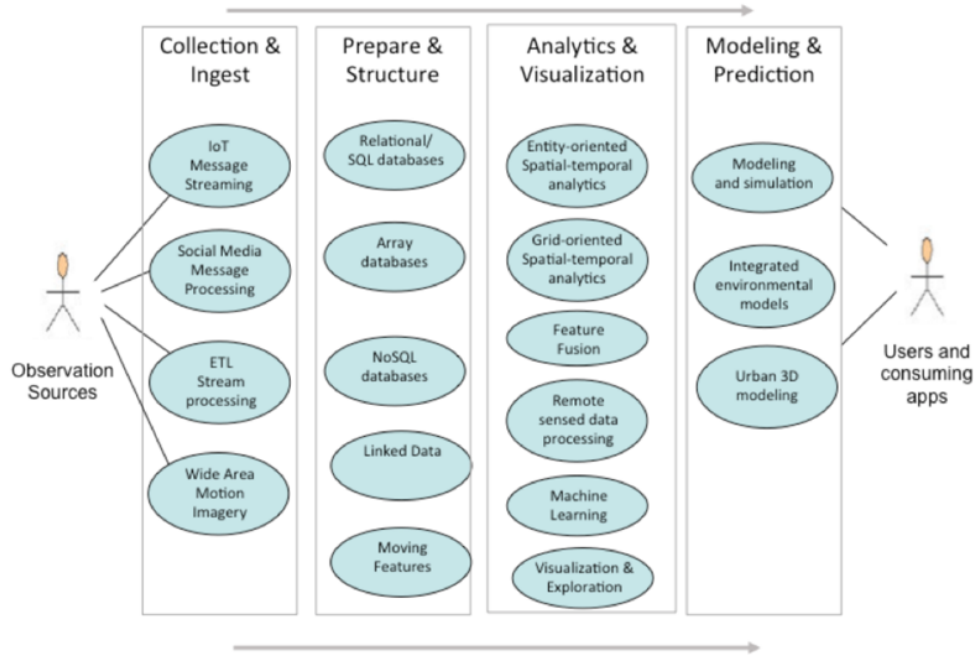


Figure 7. Big Geo Data Use Cases for Data Science

3.4. Recommendations

This Clause motivates several recommendations.

- Plan for the continued growth of Big Geo Data
- Continue to work with broad Big Data Stack to make geospatial data a routine data type for the broadest communities and to make the Big Data Stack extensible to complex analysis based on spatial temporal analytics.
- Identify common geospatial Data Science Use cases that can be reused across applications
- Promote geospatial big data developments in the Geospatial Track of ApacheCon. The geospatial track is chaired by OGC.

These recommendations are offered for uptake in OGC's Big Data Domain Working Group.

Chapter 4. Data: Data Scientists, Teams, Process

It is obvious but important to state that data management and processes are core topic of data science. For data science to be effective, data scientists needs to work in multi-disciplinary teams with an agile process.

This Clause addresses topics related to people and process for data science.

- Data Scientists and multi-disciplinary teams
- The role of tools: human augmentation
- Data Science Process
- Training and Institutionalization
- Recommendations

4.1. Data Scientists and Multi-Disciplinary Teams

To be effective, Data Scientists must work with multi-disciplinary teams. The teams need to include individuals expert in the domain of study or application. Data Scientists cannot be effective by applying generic data science tools without tuning, interpretation, and guidance from a team that provides broad understanding of the domain area. Context of the domain is needed in order that the tools are used accurately.

Several quotes are representative of the LP_DS discussions:

- “I teach my data scientists how to work on interdisciplinary teams” – Jeanne Holm, UCLA
- “I want them to be respectful and understanding of the scope that storytellers bring, what geospatial experts bring, what policymakers bring”
- “Going in as a data-scientist-with-the-answers was often counterproductive” - Regan Smyth, NatureServe
- “Running models but not understanding the drivers and whether data science can help can be misleading” - discussion group
- "Data scientists don’t learn to deploy and scale" - Underlining that their expertise alone is insufficient to solve all challenges facing companies and researchers today (such as cloud engineering).

The role of geospatial experts in the team was discussed as represented by these quotes:

- “Geospatial analysts are domain experts; data scientists tend to sit more horizontal” – Devaki Raj, CrowdAI
- “You don’t see data scientists necessarily learning geospatial technology but you expect you spatial technologists to learn data science” - discussion group
- “Data science tools become more flexible for people that have domain expertise” - discussion

group.

4.2. The Role of Tools: Human Augmentation

Without tools, data science would not be possible. But tools and the results they produce without human interpretation are not useful, or worse, they can be misleading. Tools that augment human intelligence are the most effective.

Several quotes are representative of these LP_DS discussions:

- “We can’t tensorflow our way out of this problem” – Andy Brooks
- “The r-squared is really high, but its garbage” – discussion group
- “Opportunity for commercial vendors to implement for it to become routine. User interface need not change with new stuff going on under the hood” - Marc Armstrong
- “SWAT team of nerds” - Megan Furman

4.3. Data Science Process

Several approaches to Data Science process definition or methodology were presented and discussed at LP_DS.

Stephanie Shipp, Univ of Virginia, presented their Data Science Framework. The UVA framework emphasizes working with the project sponsors to identify the problem as defined by the sponsors. Beginning with those discussions sharpens the focus; along with looking at the literature and talking to experts. Then the data discovery aspect of the framework is preconditioned by the problem identification. You don’t just start with the data that’s readily accessible. Then the process is like other data science frameworks where the data wrangling, profiling the data to assess your data quality. This is iterative work as data wrangling takes about 80% of your time. Reducing that load would leave more time for statistical modeling and analyses.

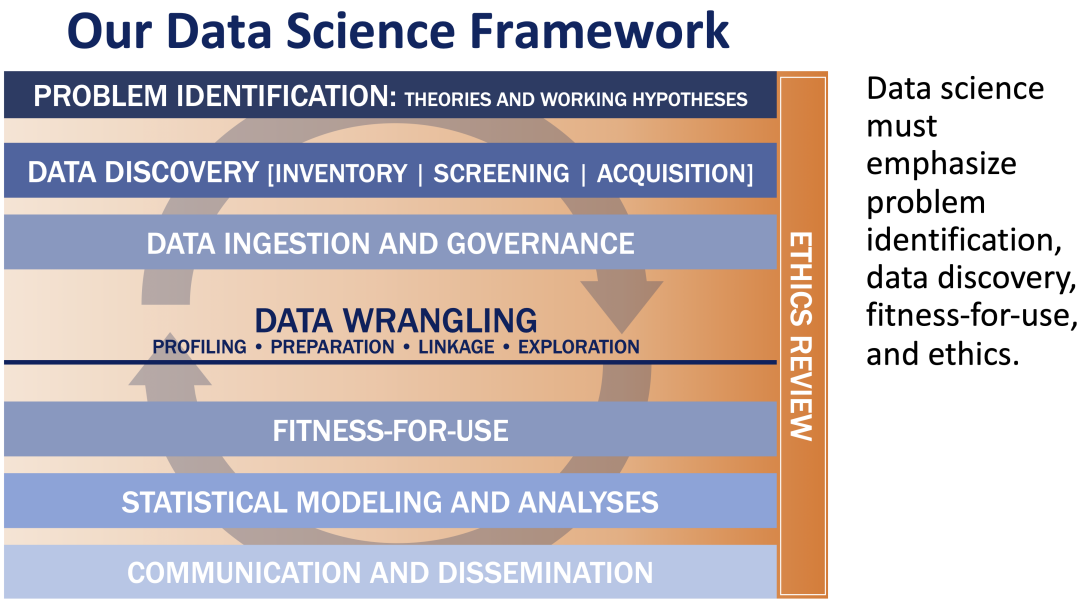


Figure 8. University of Viginia Data Science Framework

Andy Brooks, NGA Chief Scientist, used the Stanford D-School process to explain his teams approach to the data science process. The numbers in the figure is the approach that Andy used to explain their data science methodology.

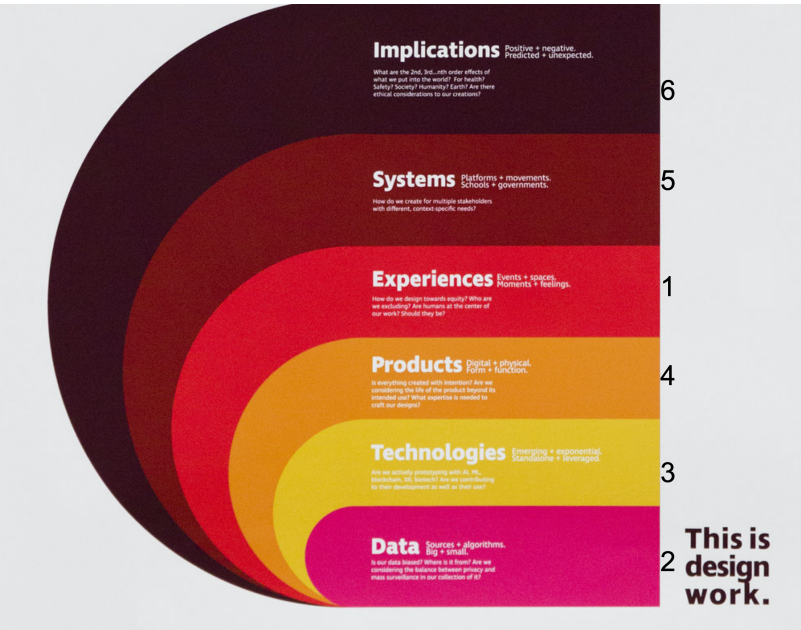


Figure 9. NGA Data Science approach based on Stanfor d-school

1. Experiences: Discussion of how do you do this today, what is the role of data in this experience, what works doesn't work; to get that ground level visceral learned experience from the people who need the results of data science.
2. Data: then start talking about the data. Where is this data coming from; how are you using it; is it "big data"; is it small data living in spreadsheets, etc. To get that sense of what that work is and how they do it. We're purposely not looking at technologies or products.
3. Technologies: Discuss what technology is used now; what do they think technology is; how is it used. Not about products, but more about the fundamental technology underneath and what's their literacy with using different forms of technology.
4. Products: Move to how does the data, technologies and products all roll together in an experience of what they're trying to do. First understanding fundamental things like the data is not really that good; or the underlying technology doesn't work; or the policy isn't enabling them. That's why products come along later in the process.
5. Systems: Further along come discussions on how to understand the system; to scale what you're trying to do; who are those people that you need to get to those teams you need to work with across the organization.
6. Implications: then consider the implications, e.g., of speeding up a workflow and making it that much faster, because there is that thing where it's like well it used to take ten people two weeks to do one thing that would spit something out and now it takes like one person clicking on a script and they can do it in like ten minutes well like there's a lot of implications for that.

The Azure Data Science Lifecycle was not presented in the LP_DS Summit, but as it is consistent with the discussions at the LP_DS Summit is presented here. It is an iterative data science methodology that focuses on team collaboration and learning; with an initial business understanding prior to data acquisition and modeling. It is a generic process that can be

implemented with a variety of tools.

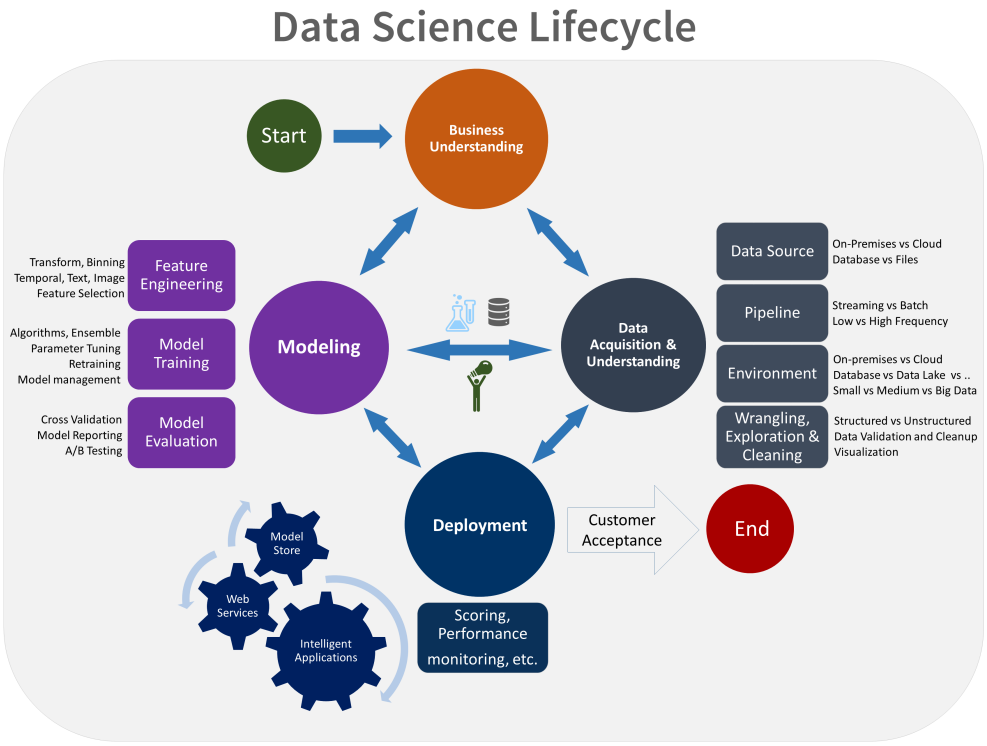


Figure 10. Azure Data Science Lifecycle

4.4. Training and Institutionalization

To support institutionalization of geospatial data science, data scientists need training and education and organizations need to persist the best practices and standards that emerge from successful projects.

- The Geospatial Software Institute concept study as presented at LP_DS by Anand Padmanabhan, University of Illinois, was a US National Science Foundation sponsored study to conceptualize GSI as a long-term hub of excellence in geospatial software to serve diverse research and education communities. The CyberGIS center at UIUC led the conceptualization project that include key aspects of geospatial data science. The CyberGIS Summer school provides education and training for indviduals learning geospatial data science.
- The Data Science Foundation as presented at LP_DS by Jeanne Holm, City of Los Angeles, is a partnership between The City of LA with local colleges and universities; as a resource in data science and data-driven decision making for City Government.
- The NGA Data Corps is a targeted initiative to support Data Scientists in solving complex, high-stakes data problems; teaching data skills to colleagues; and education to ensure keeping pace with the latest tools, techniques and technology.
- The Defense Digital Service looks to both attract and create talent over time to meld tricky domain expertise with data science training. Megan Furman, DDS, spoke of her team as a "Swat Team of Nerds."
- AIST Artificial Intelligence Research Center (AIRC) as presented at LP_DS by Satoshi Sekiguchi, offers a “venue” for open innovation that connects the proprietary data and expertise in machine learning, simulation technology, natural language processing, and development of

computational architecture for AI. AIRC is a public organization that coordinates AI technology by promoting the sharing of data that cannot be made public by businesses and universities.

- ESA Φ-lab is part of the ESA Earth Observation (EO) Programme's Φ-Department developing future systems for earth observation. Φ-lab convenes data scientists and technologists from across the World to develop research agendas on the relevance for EO of emerging technology topics including AI, distributed ledgers and quantum computing. The Φ-lab's mission is to accelerate the future of earth observation, by helping Europe's earth observation and space researchers and companies adopt disruptive technologies and methods.
- The Open Geospatial Consortium provides the processes for communities to advance geospatial data science. The OGC Geospatial AI Domain Working Group and the OGC Big Data Domain Working Group are chartered to foster discussion, to identify community best practices and as needed initiate standardization relevant to geospatial data science.

4.5. Recommendations

This preceding sections of this Clause motivate several recommendations.

- Identify and promote Community Practices and Best Practices for Geospatial Data Science.
- Promote the development of institutes that capture current practices, research advancements and training practitioners in geospatial data science.

Chapter 5. Data: Data Management

This Clause addresses topics related to people and process for data science.

- Data Management Policies
- Data Base Management Systems
- Cloud Native Data and Processing
- Geospatial Data Platforms and Ecosystems
- Recommendations

5.1. Data Management Policies

Accessibility to data involves many issues. This section addresses the policies associated with making data available.

Several LP_DS discussions described the desirable situation where a data is accessible with known license, provenance and quality. There are trends to make research datasets more readily available due to requirements coming from the sponsors of the research, e.g., US National Science Foundation. Some of the discussions suggested that 3rd party organizations take the responsibility of a clearing house to make the data well documented and accessible without relying on the research team. Others suggested that tools like GitHub and Jupyter Notebooks are making easy for researchers to publish their results.

Quotes from LP_DS regarding data agreements:

- "How can we go beyond the current state which is a big lake of data where you just have to look for it by yourself and figuring out yourself if the data is good or not?"
- "We need to address our data sharing agreements they often get in the way and hold up our work. There’s ways to overcome that that are not terribly hard. "
- "A role for entities that behave as data authorities in the future that server as centralized sources of data; they ensure that the data and the metadata is good quality and follows the standards and they clarify data sharing agreements."
- "Need better methods to make data communicate the range of data agreements where not everything is fully open or is open under certain conditions."
- "Would be helpful to extend use schema.org and others to include the data sharing agreement aspect, so for example, you could search for images that have certain copyright or open new standards"

Quotes from LP_DS regarding dataset hosting and clearinghouses:

- "In cases where researchers received federal grant money and they are supposed to publish data openly, but the challenge that you briefly mention is the incentive and resources that they don’t have so they do the research everything ready results are published in a paper at the end"
- "The PI and the student or postdoc don’t have the incentives to put it in a public repo or documented or don’t have the technical resources. Usually they’re more of a scientist than an

engineer. I think one thing that would be helpful is policies and resources in those grants to help grantees to share their data openly beyond just a policy but on the implementation side"

- "I think the private sector needs to step up and make data accessibility and data quality an important issue and be able to be part of this research community in terms of providing good high quality data for very important issues."
- "For federal data, data.gov serves the purpose of a clearinghouse."
- "We are at Google so let's give a shout out to Google dataset search. You can now specifically search for datasets."

Quotes from LP_DS regarding actions to take to further increase and ensure availability of open data and make recommendations to them

- "While open data has soared in terms of quality and quantity but making it more standardized if possible as it comes in all different formats and also needs to come with more metadata as well. It's much better than it was but it still has a way to go in terms of being standardized; another place for standards"
- "Big open archives of observation data are readily available but what's lacking is the availability of in-situ data, reference data and training data. Need to motivate research groups to reprocess their reference data and bring it into the standardized format."
- "When we started building Data.Gov, I had to convince folks at 175 federal agencies that it was okay to publish messy data; it was okay if your data was incomplete; it was okay to publish the data the best way you could."
- "We can't ignore the political issues around keeping data that may be inconvenient for one administration or another. Keeping data open over time because people are running businesses and they are running in analyses and data journalists are linking to this data all the time and they need to have real-time accurate data that is standardized to the extent possible but accessible over time"

Other issues about data sharing policies included discussion of data privacy and data ethics. This will be take up on Clause 10 on the emerging trend of Data Ethics.

Access to training data sets for machine learning is discussed in Clause 7.

While not explicitly mentioned, the FAIR Principles capture much of the discussion about data management policies during LP_DS. A recent issue of the Data Intelligence journal addressed the history, progress and possible futures around the FAIR Principles. OGC lists the FAIR Principles in its mission statement.

Working with the Research Data Alliance to develop data management policy topics should be considered

5.2. Data Management Strategies and Query Languages

Strategies for data base management continue to change. Data strategies for modern applications require many different data types and workloads. Data types include relational, document, spatial, graph/linked data. Workloads include transactions, analytics, machine learning, streaming, etc.

Each data type and workload requires different data strategies. Jayant Sharma, Oracle, described two approaches to data strategy: single-purpose "best-of-breed" for each data type and workload; or a converged database to address all data types and workloads. A converged database does not mean centralized database. The benefits of a converged database is that it supports mixing of workloads, data types and algorithms; and it prevents data fragmentation; and enables powerful synergies across features, e.g., SQL and transactions across any data type.

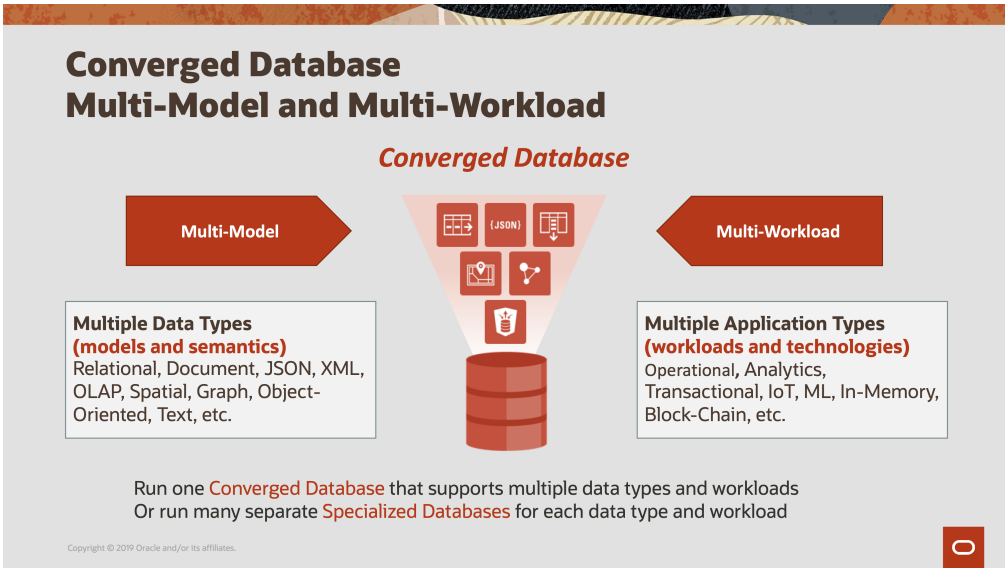


Figure 11. Coverage Database Data Strategy

The LP_DS included discussion of SQL and GQL by Keith Hare, Convener of ISO/IEC JTC1 SC32 WG 3, that manages standards for data languages. The SQL standards include support for geospatial data with spatial queries, temporal queries, and very recently multi-dimensional arrays. Currently SC32 WG3 is developing standards to address property graphs both extensions to SQL for property graph queries as well as a declarative property graph query language titled GQL. OGC is contributing to the spatial capabilities of GQL.

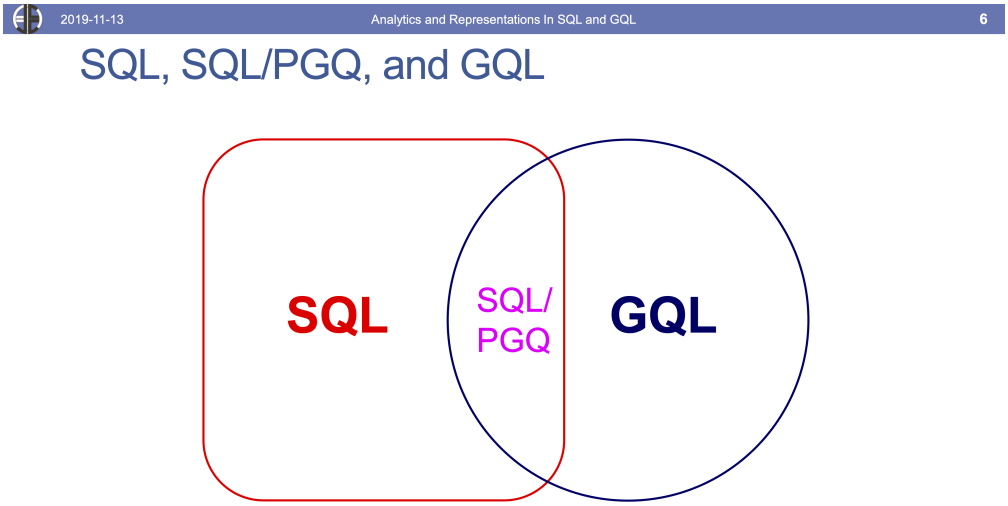


Figure 12. Property Graph Query Languages and SQL

5.3. Cloud Native Data and Processing

As described in Clause 3, an innovative stack of software and interfaces was developed to address

big data. The Big Data Platform was grew with the development of Cloud Computing. Several discussions in LP_DS discussed advantages of a "cloud-native" strategy for handling big geo data. Mark Korver, Amazon AWS, described the growth in size and functionality provided by cloud computing. AWS hosts over two trillion objects. Object stores offer not only data storage but the bigger advantage is having the data and computing in close proximity. Satoshi Sekiguchi, AIST, spoke about the key features of the cloud native strategy for their AI Bridging Cloud Infrastructure (ABCI) system and how it supports the sharing, distribution and operation of AI and Machine Learning on large data stores.

To support Cloud-Native strategy several specification activities were discussed as needing standardization activities: Spatio-Temporal Asset Catalog (STAC), Cloud-optimized GeoTIFF, HDF for the cloud, ZARR and X Array, and OGC APIs.

OGC APIs providing access to cloud hosted data and anaytics are under development. The first standard released in 2019 was OGC API Features - Core. STAC is a consistent with OGC API Features as they were developed in several sprints where the two specs were worked concurrently. For coverages including raster and other data structures

Lauren Bennett, Esri, conclude one of the discussions on this topic with: "I guess from from my perspective it’s about making it really easy for people to go between these different platforms; to bring together these diverse sets of data; and a diverse set of methods, models, algorithms; that are coming from all over the place. The standards allow integration that’s crucial as people we need people not to be stuck in one place."

5.4. Geospatial Data Platforms and Ecosystems

Multiple organizations are defining platforms for geospatial data science on top of cloud computing and data stores. ESA’s EO Platform Ecosystem and several other approaches have been discussed in the OGC EO Exploitation Platform Domain Working Group. Emergence of community practices to be used in common is an objective of the working group.

Patrick Griffiths, ESA, presented an EO Platform Ecosystem which depicts the earth observation platform ecosystem. The Platform builds on distributed storage environments co-located with distributed compute environments. Top two layers are the exploitation tier consisting of a platform services layer as well as an exploitation layer. The Platform services layer includes data cubes and analytic environments to take advantage of all of this data. There is much discussion about "data cubes", but there is agreement on the need to access the data at the per-pixel level without having to worry about pre-processing and on the need for data management which often took up to 70 or 80 percent of the earth observation scientist’s time.

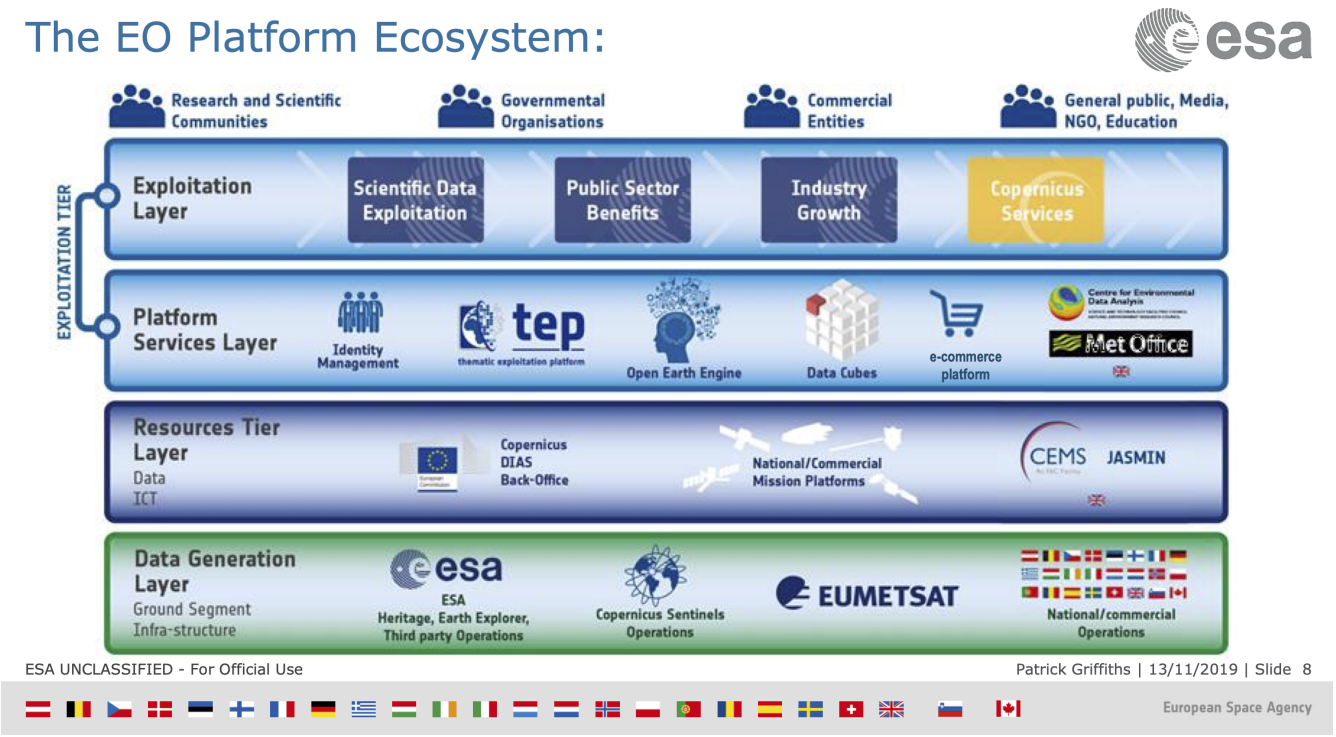


Figure 13. EO Platform Ecosystem

5.5. Recommendations

- Define OGC Community Practices for data sharing agreements including how to find data sets based on agreement of interest.
- Apply the FAIR principles to data management policies issues identified for geospatial data.
- OGC to support GQL development, e.g., by providing geospatial use cases and sample queries for linked spatial data.
- Advance a slate of cloud-native standards for geospatial data.
- Develop and publish an OGC Community Practice for Geospatial Coverage Data Cubes.
- Develop and publish an OGC Community Practice for EO Exploitation Platform.

Chapter 6. Tools: Geospatial Representations and Analytics

Geospatial, analytics, and data science workflows are converging. This first section in the tools category of the white paper considers several "sovled problems" applied in data rich environment. Geospatial representations and analytics are well established. This clause highlights the value that mature geospatial methods uniquely bring to data science. The topics in this clause form a substantial basis for what geospatial data and processing brings to the data science world.

```
"Analytics, data science and location intelligence are converging, at scale, in real
time, leading to new challenges you can solve for with these capabilities"
Tod Mostak, OmniSci (formerly MapD)
```

This Clause addresses these topics:

- Maps: cartographic display of data
- Statistical Geography
- Space-Time Analytics
- CyberGIS
- Scientific Computing: Notebooks, Python, R
- GPU-accelerated Geo Analytics
- Recommendations

6.1. Maps: cartographic display of data

Cartography is the study and practice of making maps. Combining science, aesthetics, and technique, cartography builds on the premise that reality can be modeled in ways that communicate spatial information effectively.

Nearly every presentation in the LP_DS summit contained maps. Not every map is a study in data science, but maps are a powerful tool for the visual display of data. Making maps is an art and science that can tell powerful stories, that can be both accurate or deceptive.

For example, this white paper is being prepared during the outbreak of the corona virus outbreak that began in Wuhan China and may be come a global pandemic. While not part of the LP_DS Summit, Kenneth Field - the current Chair of the ICA Map Design Commission - blogged on need for Responsibly Mapping Coronavirus. Below are two maps displaying the same data. The first map is a emotive red color, of total cases per province. Hubei Province is a massive outlier, a really massive outlier that the map above doesn't properly reflect. The second map tells a more accurate story.

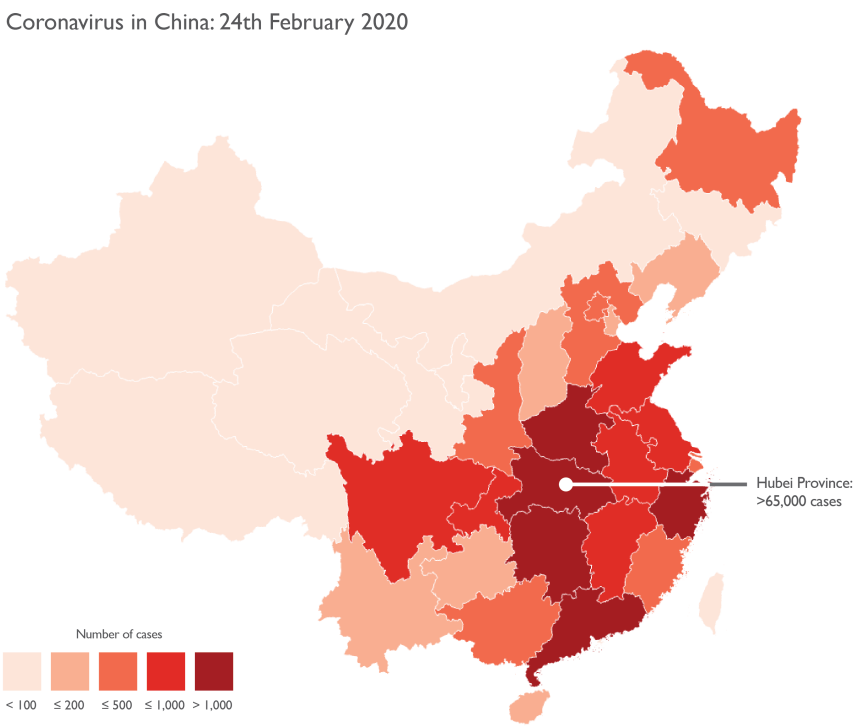


Figure 14. Corono virus map: red choropleth

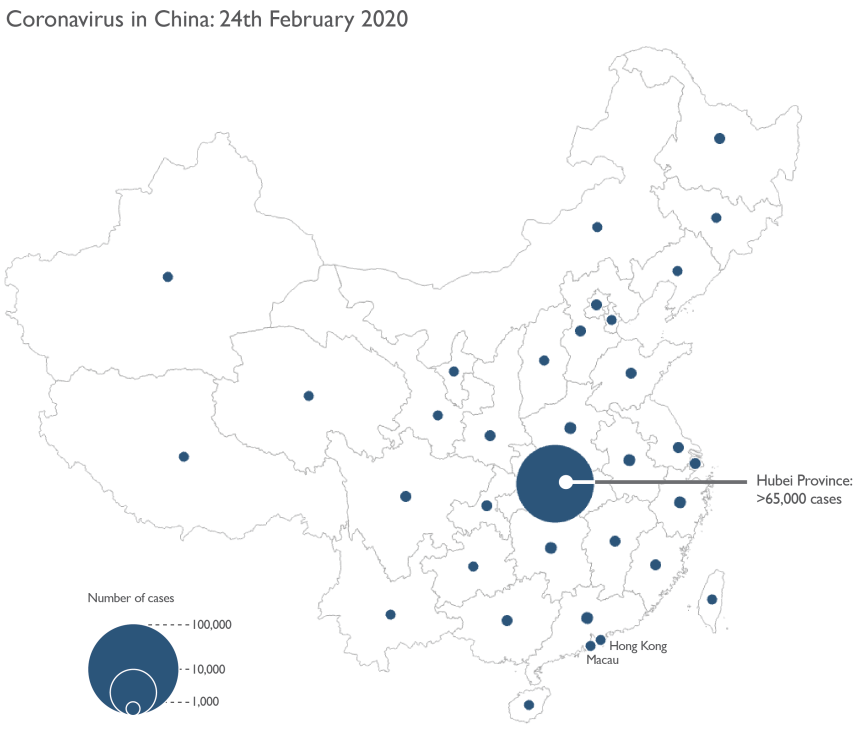


Figure 15. Corono virus map: blue proportional symbol

shows

Field concludes his blog with: "The key to informing is to work with the data and to not imbue it with misguided or sensationalist data processing or symbology" and "pick a technique that supports

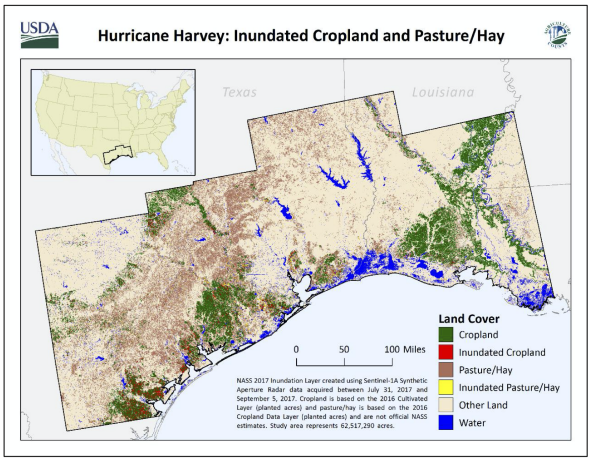
the telling of that story, process the data and choose symbols that are suitable, and avoid making a map that misguides, misinforms."

Accurately telling stories and conveying experiences with data using maps is a key contribution of geospatial sciences to data science.

6.2. Statistical Geography

Statistical geography is the study and practice of collecting, analysing and presenting data that has a geographic or areal dimension, such as census or demographics data. It uses techniques from spatial analysis, but also encompasses geographical activities such as the defining and naming of geographical regions for statistical purposes.

Wendy Martinez, President of the American Statistical Association presented several examples of statistical geography at the LP_DS Summit. The one shown here is a calculation of inundation of crop area due to Hurricane Harvey. The figure show both a map and a table. The map is effective in telling the geographic visualization. The table is based a geographic analysis based on the fusion of several data layers and the areal calculation function resulting in the precent of inundated land. This is a relatively simple example of geospatial calculation of statistical data.



Crop Type	Percent Inundated
Corn	14.54%
Cotton	14.53%
Fallow/Idle Cropland	9.47%
Oats	10.39%
Rice	7.43%
Sorghum	25.72%
Winter Wheat	11.45%
Total Cropland	10.16%
Pasture/Hay	3.68%

Total Area Analyzed
Total: 62,517,290 acres
Cropland: 7,061,403 acres
Pasture/Hay: 9,448,350 acres

Figure 16. Inundation Map and Analysis by USDA

The methods of data science can be seen as emerging from statistics [David Donoho]. The combination of statistics with geospatial science is powerful tool. The US Federal Committee on Statistical Methodology includes a Geospatial Interest Group which coordinates methodological information related to geospatial data across federal agencies. The FCSM GIG recently held a workshop that considered the question: What are the unique aspects of geospatial data to consider when determining data quality in the context of integrated data products? Key topics were geospatial representation, error propagation models, geometry and spatial relationships. Critical issues to be considered when modeling and integrating geospatial data include:

- Geometry of geospatial objects is scale-dependent
  - What level of geometry at different scales?
  - How can we integrate data a different scales/resolutions?

- Spatial relationships between objects
  - Overlap or inclusion
  - Direction
  - Distance

Wendy Martinez concluded her presentation at the LP_DS summit on the topic of data ethics. Ethics will be addressed in Clause 10 of this white paper.

The OGC Statistical Domain Working Group is chartered to identify requirements and use cases of how geospatial and statistical standards can support the integration of geospatial information into the statistical system and for the purposes of broad discovery, analysis and use. The OGC Discrete Global Grid System standard has been identified by the Statistical DWG as a standard to considered in their work.

6.3. Space-Time Analytics

Temporal analytics is an area of excellent research that is becoming increasingly more important and impactful with the availability of data and big data processing.

Kathleen Stewart, UMCP, presentation at LP_DS Summit provides an excellent example of the discussions about temporal analytics during the summit. Early in the summit there was reference to Kathleen's publication Computation and Visualization for Understanding Dynamics in Geographic Domains. Her presentation focused on "New opportunities through big mobility data analytics." Space-time patterns are available from different data sources: GPS waypoint data, cell phone data, location-based app data, as well as other sensor, e.g., fitness trackers.

In the figure below we see big trajectory data (GPS waypoints transformed into trajectories) useful for highlighting travel behaviors of different groups. We want to expose different dynamic behaviors over space and time. Differences are observed in road transportation for differing vehicles, eg., urban-rural differences. This analysis shows different patterns in those settings which may have differences for risk exposure and planning for major mass evacuation whether it's for flooding or wildfires.

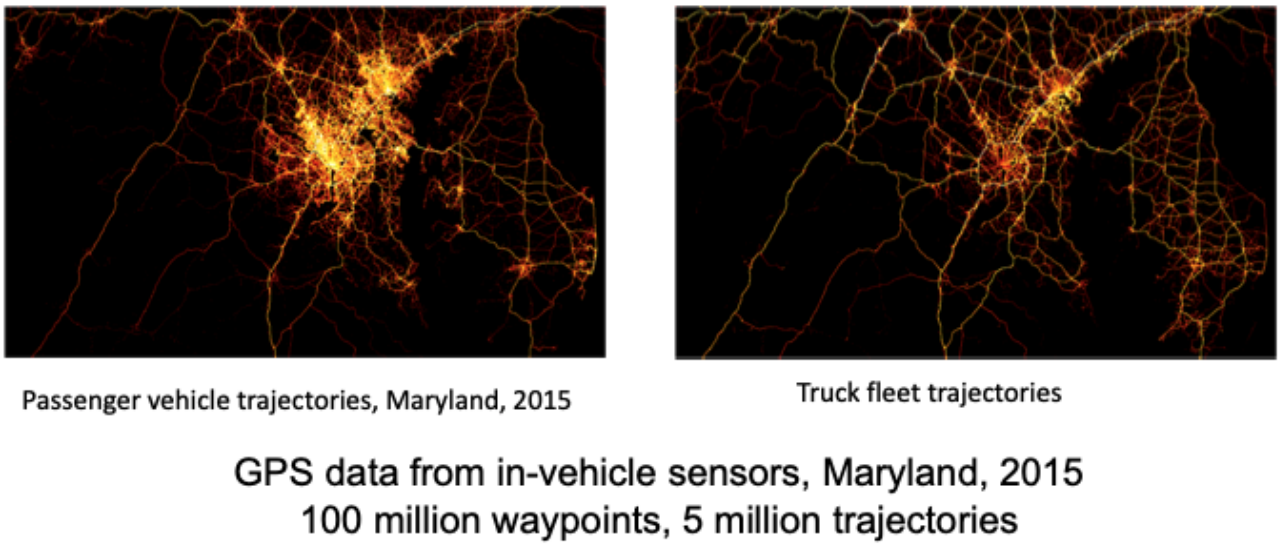


Figure 17. Space-time trajectories for different vehicles

Spatial-temporal analytics requires trajectory reconstruction algorithms. This can involve: snap way-points of a trip to road segments; and filling segment gaps by heuristic algorithms.

The OGC Moving Features Standards Woring Group considers applications using moving feature data, typically on vehicles and pedestrians. Innovative applications are expected to require the overlay and integration of moving feature data from different sources to create more social and business values. Efforts in this direction are encouraged by ensuring smoother data exchange because handling and integrating moving feature data will broaden the market for geo-spatial information such as Geospatial Big Data Analysis. The Moving Features SWG has created a suite of OGC Moving Features Standards

In Clause 10, the importance of spatial-temporal analysis is discussed in the context of data streaming or "Fast Data."

```
"We finally have the data and computing that we didn't have those many years ago when
the research was being conducted. We were talking about trajectory-like objects many
years ago and now we have them for real. We are discovering new things that we didn't
really think about before because we didn't understand how the technology was going to
deliver these things" - Kathleen Stewart
```

6.4. CyberGIS

CyberGIS is a Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. Cyberinfrastructure (CI) integrates distributed information and communication technologies for coordinated knowledge discovery. In the linked article, Shaowen Wang describes how CyberGIS provides a framework for the synthesis of CI, geographic information systems (GIS), and spatial analysis (broadly including spatial modeling). The framework focuses on enabling computationally intensive and collaborative geographic problem solving.

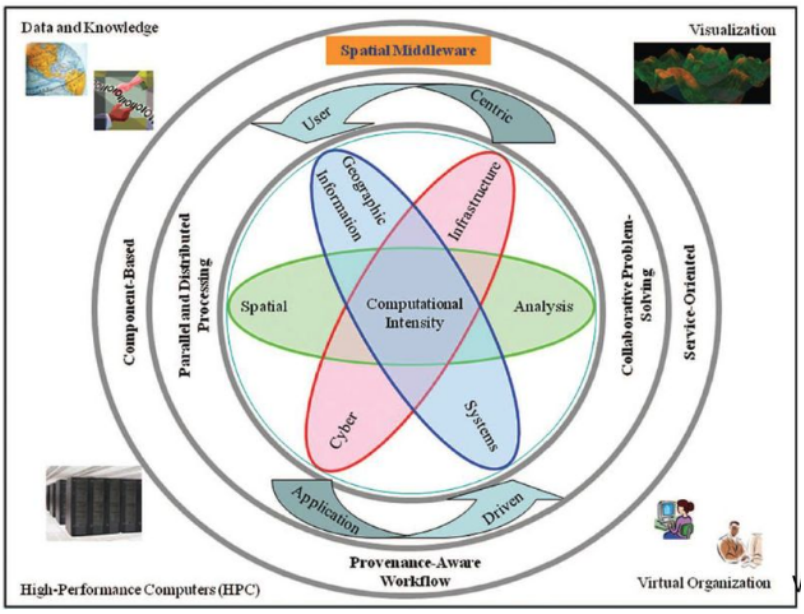


Figure 18. CyberGIS Framework

During the LP_DS Summit, Anand Padmanabhan, University of Illinois, describe applications of the CyberGIS Framework. The CyberGIS approach enabled creation of a flood inundataion map at continental scale. Hydrologists seeking to address flood mapping on a national scale that had not been done before as it required a scale of computation not previously available. The US National National Hydrography data sets from USGS were used in calculation of Height Above the Nearest Drainage (HAND) based on terrain models. For more information: A CyberGIS Approach to Generating High-resolution Height Above Nearest Drainage (HAND) Raster for National Flood Mapping

Anand also presented about CyberGIS-Jupyter for handling big data and analysis at scale and make this results sharable and reproducible. CyberGIS-Jupyter project extends the CyberGIS framework for achieving data-intensive, reproducible, and scalable geospatial analytics using Jupyter Notebook. The framework adapts the Notebook with built-in cyberGIS capabilities to accelerate gateway application development and sharing while associated data, analytics, and workflow runtime environments are encapsulated into application packages that can be elastically reproduced through cloud computing approaches. As a desirable outcome, data-intensive and scalable geospatial analytics can be efficiently developed and improved, and seamlessly reproduced among multidisciplinary users in a novel cyberGIS science gateway environment.

6.5. Scientific Computing: Notebooks, Python, R

Recent advances in scientific computing have being used to deal with big geo data and to advance geospatial data science. These include the use of notebooks, e.g., Jupyter Notebooks, as well as languages well suited to data analytics such as Python and R.

Jupyter Notebooks have rapidly become a popular method for sharing analysis approaches, linkage to datasets and computation resources in a cloud friendly fashion. Multiple presentations in LP_DS Summit described use of Jupyter Notebooks. The figure presented by Jay Theodore, Esri, shows how notebooks serve as a container for workflow with links to the computing resources.

Figure 19. Notebooks in Esri ArcGIS

Python is a modern, general-purpose, object-oriented, high-level programming language. Python has a strong position in scientific computing with a large community of users, easy to find help and documentation. There is an extensive ecosystem of scientific libraries and environments: numpy for Numerical Python, scipy for Scientific Python, matplotlib a graphics library. It has great performance due to close integration with time-tested and highly optimized codes written in C and Fortran. Python also refers to the standard implementation of an interpreter (cython). The most common way to use the Python programming language is to use the Python interpreter to run python code.

Extensions of Python for geospatial computation are available as described here and there. Many OGC members have developed their own extensions to Python.

The R programming language and environment supports statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. There are numerous resources for introduction to Python data science "ecosystem", e.g., Python Data Science Handbook is quite comprehensive.

Extensions of R for geospatial computation are available as described here and there.

OGC Testbed 16 will be addressing Earth Observation Application Packages with Jupyter Notebooks.

OGC APIs are being developed independent of programming languages but intended to be compatible with taking advantage of Python and R.

The adopted OGC GeoAPI standard provides a Java API. A Python profile of GeoAPI is under development.

6.6. GPU-accelerated Geo Analytics

GPU based computing is improving the performance of many of the topics listed earlier in this Clause. (Heterogeneous computing beyond GPUs is address in Clause 10.)

Tod Mostak, OmniSci, performed a demonstration in LP_DS Summit that showed the intersection of analytics and data science based on the GPU accelerated calculations. The demo included visualize aircraft flight tracks as 5 billion points. The calculations included spatial bins in seconds and pivoting this table to generate a huge SQL query behind the scenes. The GPU accelerated k-means algorithm which takes only seconds. The points were then then clustered and visualized as trajectories.

The OGC Community Standards for 3D Visualization 3DTiles and i3S make use of GPU accelerated visualization through use of the Khronos Group GL Transmission Format (glTF). glTF is an efficient, extensible, interoperable format for the transmission and loading of 3D content. glTV was developed to mirror the GPU APIs.

Milind Naphade, described how NVIDIA created cuSpatial: a free library for GPU acceleration of common spatial operations as listed in the figure. The acceleration provides the instantaneous results of hypothesis testing, e.g., clustering, bu several orders of magnitude acceleration.

Layer	0.10/0.11 Functionality	Functionality Roadmap (2020)
High-level Analytics	C++ Library w. Python bindings enabling distance, speed, trajectory similarity, trajectory clustering	C++ Library w. Python bindings for additional spatio-temporal trajectory clustering, acceleration, dwell-time, salient locations, trajectory anomaly detection, origin destination, etc.
Graph layer	cuGraph	Map matching, Djikstra algorithm, Routing
Query layer	Nearest Neighbor, Range Search	KNN, Spatiotemporal range search and joins
Index layer	Grid, Quad Tree	R-Tree, Geohash, Voronoi Tessellation
Geo-operations	Point in polygon (PIP), Haversine distance, Hausdorff distance, lat-lon to xy transformation	Line intersecting polygon, Other distance functions, Polygon intersection, union
Geo-representation	Shape primitives, points, polylines, polygons	Additional shape primitives

Figure 20. cuSpatial - GPU Acceleration of common spatial processing functions

6.7. Recommendations

Recommendations for consideration by the OGC Big Data Domain Working Group:

- Promote development of Big Data Stack approaches for Spatial-temporal analytics and Streaming analytics
- Promote development of OGC Community Practices for for geospatial cyberinfrastructure, e.g., CyberGIS
- Promote discussion and development of computing using Notebooks, R and Python-oriented APIs, e.g., results from OGC Innovation Program initiatives.

Recommendations for consideration by the Moving Features SWG and Temporal DWG:

- Promote development of OGC Community Practices for spatial-temporal analytics.
- Propose use cases for Edge Computing: temporal analysis of streaming data

Recommendations for consideration by the Statistical DWG

- Promote development of OGC Community Practices for geospatial data science based on Statistical Geography
- Promote discussion of the Impact of big data platforms analytics on statistical geography

The OGC Testbed 16 results regarding notebooks and python-oriented APIs should be considered by OGC, e.g., the OGC Big Data DWG and Earth Observation Exploitation DWG.

Chapter 7. Tools: AI and Machine Learning for Geospatial

This clause addresses how recent advances in Artificial Intelligence, in particular Machine Learning, have been applied to geospatial data. The application of machine learning to geospatial data is described here as a significant advance in geospatial data science.

This Clause contains these sections:

- The emergence of Artificial Intelligence
- Challenges with AI
- Training Sets and Benchmarks
- Augmenting Machine Learning
- Look forward to the future of AI
- Recommendations

7.1. The third emergence of Artificial Intelligence

Artificial Intelligence is in its third generation of technology development. In the LP_DS Summit, Satoshi Sekiguchi, AIST, presented a figure showing these three generations: where the top half of the diagram shows development of AI as rule-based, symbolic manipulation with inference engines; and the bottom half address data driven AI based on pattern recognition and machine learning. In this clasue the emphasis will be on Machine Learning. Knowledge Graphs as a form of AI will be considered in Clause 8.

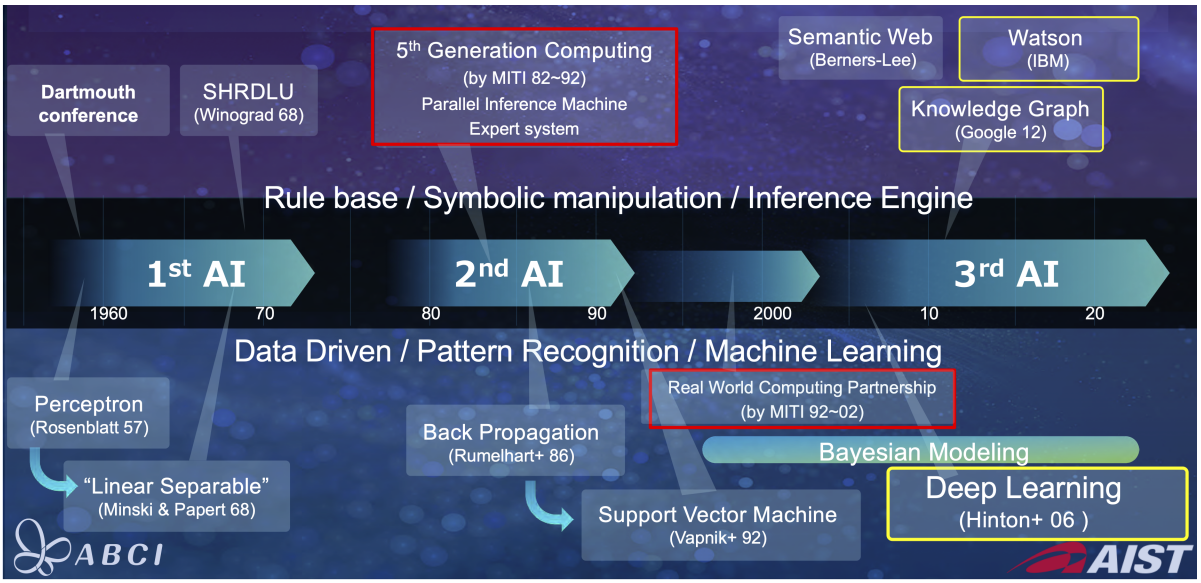


Figure 21. Generations of Artificial Intelligence development

The recent advances in machine learning has been driven in large part by the ready availability of training data commonly based on computer vision benchmark datasets such as Imagenet etc. Satoshi Sekiguchi portrayed this as an advancement in Data Science where the New Oil (big data) meets the New Engine (deep learning with high performance computing and big data computing).

Within HPC, Satoshi considers traditional HPC applications including mathematical or physical models. The mixing of augmenting machine learning training based on data with hypothesis and models will be address later in this clause.

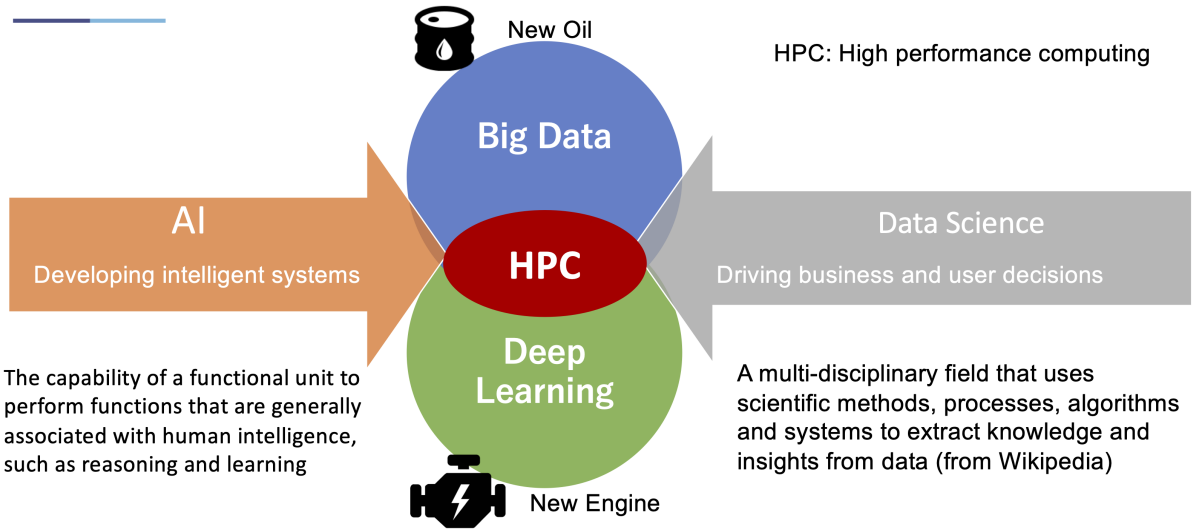


Figure 22. AI/Machine Learning and Data Science

Several discussions about AI and Machine Learning are captured in these quotes

- "CEOs of the big companies 40% of the top performers think that AI is going to be a game-changer for the industry; bigger than cloud, mobile, IoT, blockchain or APIs" - Philippe Cases
- "AI is a social construct more so than a set of technologies - this is based on analysis of more than a hundred thousand documents in data science and analytics" - Andre Skupin
- "Mostly what were are doing in geospatial is using machine learning" - group
- "Only machine learning can help scale to consume the exponential growth of video streams" - Nils Lahr
- "What we've been doing is trying to solve complex problems using tools; it's about trying to solve problems" - group
- "Need to consider both AI and intelligence augmentation" - group

Jay Theodore, Esri, provide an example of how Machine Learning for imagery can be deployed as shown in the Figure. The full workflow where you collect the data, prep the data you use it for training, then build and validate models, deploy them and then once you deploy them you probably want to run the inferencing at scale. This is the full end-to-end workflow that typically might happen in the cloud. Then for deploying to the edge you pick and choose what you want to do at the edge and deploy it at the edge. You might train the model at the edge to avoid sending the data to the cloud. Train the model at the edge so no privacy laws are broken by pushing data up to the cloud.

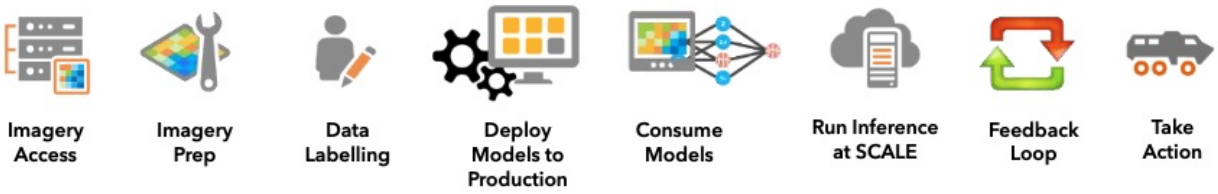


Figure 23. Machine Learning with Imagery life cycle by Esri

An example of the effectiveness of machine learning is presented here. Several more examples will be presented as applications in Clause 9. Regan Smyth, NatureServe, has applied machine learning to mapping of species habitats. The figures shown here are predicted habitats regions for a type of salamander. The figures depict the improvement in predicted suitability of the right over the left map. Furthermore, the efficiency of making these maps is greatly improved with machine learning. In the past year NatureServe has done this for 2,000 species by pulling together data collected by hundreds of people in the field building a kind of cloud-based environment where we have a team of scientists collaborating on the modeling and then using tools to get that information back out to our scientists to review it and tell us how well the models have done.

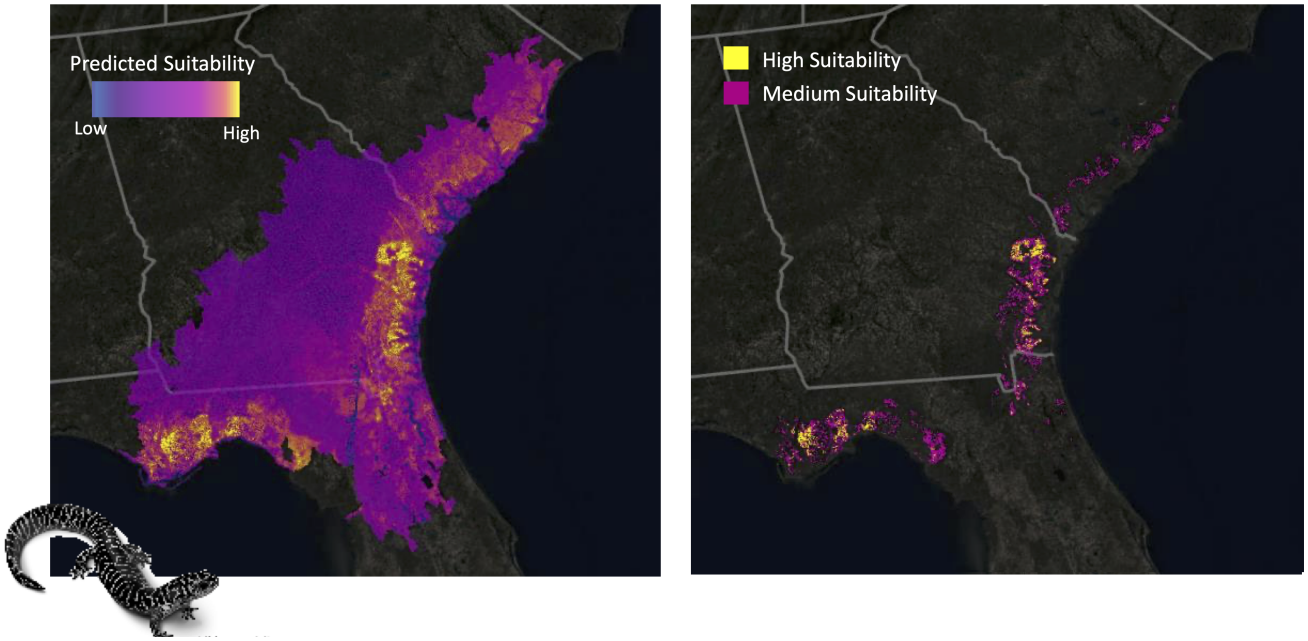


Figure 24. Effectiveness of Machine Learning in habitat identification

7.2. Challenges with AI

While there are many examples of the benefits of machine learning, there was a substantial disucssion in the LP_DS Summit about the current challenges with machine learning. The figure below was presented by Nils Lahr, depict multiple of the challenges discussed.

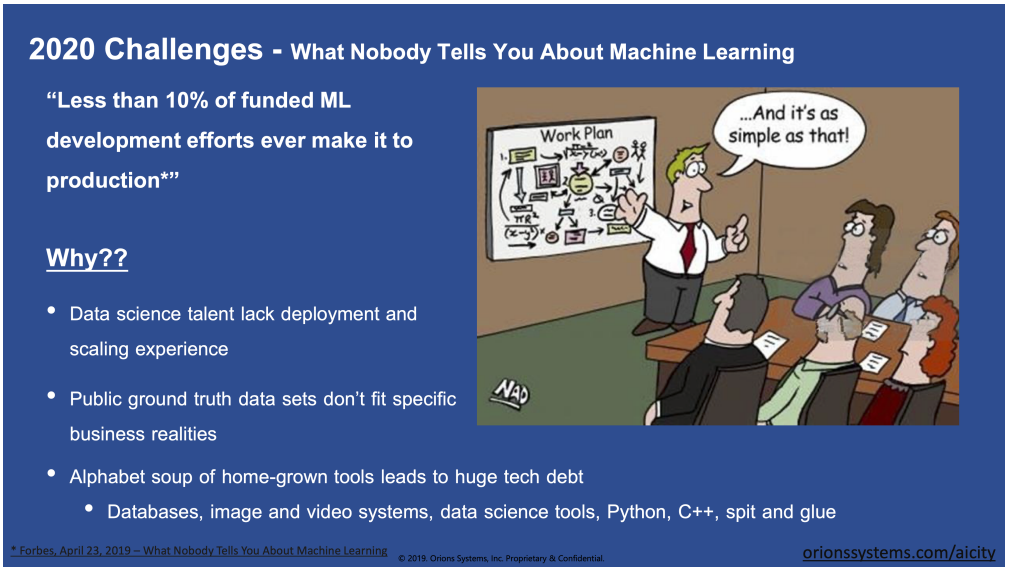


Figure 25. Current Challenges with Machine Learning

Challenges with Machine Learning are reflected in these quotes from the workshop:

- "Now there a lot of "hello world" experiments. What we need are real world solutions; pushing these experiments into real production to build trust. What's needed is the core engineering of building end-to-end systems." - Milind Naphade
- "The biggest challenge is lack of data. Some companies have the resuorces to acquire the data they need to make progress but and there are small companies along with niche applications that don't have sufficient data. Ecosystems are developing to gather enough data to gain confidence in decisions." - Anand Kannan
- "Benchmarks data sets with labels are needed to develop the systems so end users can actually start having confidence that this really works" - Milind Naphade
- "There is a lack of geospatial training data catalogs. This leads to biased or incorrect results and the inability to capture wide range of possible outcomes in space and time" - Hamed Alemohammad
- "A really good vehicle detection model in the Midwest US may look very different than a really good vehicle detection and tracking model in Shanghai or Beijing or Mumbai." - Milind Naphade, NVIDIA Metropolis
- "You can create all the models you want all day long and then all of a sudden something real happens and you realize that the models aren't what you needed" - Nils Lahr

7.3. Training Sets and Benchmarks

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Since 2010, the ImageNet project has run an annual software contest, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where software programs compete to correctly classify and detect objects and scenes. ImageNet and the annual ILSVRCs have been essential to advancement of Machine Learning. According to an article in The Economist the current excitement about the field, can be traced back to 2012 and an online contest called the ImageNet Challenge.

Several activities have geospatial data sets comparable to ImageNet:

BigEarthNet - a benchmark archive constructed by TU Berlin with European Research Council funding - consisting of 590,326 Sentinel-2 image patches from atmospherically corrected tiles acquired between June 2017 and May 2018 over the 10 European countries. Each image patch was annotated by the multiple land-cover classes (i.e., multi-labels) that were provided from the CORINE Land Cover database of the year 2018.

SpaceNet is a corpus of commercial satellite imagery and labeled training data to use for machine learning research. SpaceNet focuses on four open source key pillars: data, challenges, algorithms, and tools. SpaceNet Challenge Dataset's have a combination of very high resolution satellite imagery and high quality corresponding labels for foundational mapping features such as building footprints or road networks.

During LP_DS Summit, Hamed Alemohammad, Radiant Earth, applauded the BigEarthNet and SpaceNet activities, while also pointing out that more work is needed on training datasets and benchmarks to address problems like the lack of diversity, accessibility of data, interoperability of data sets, and the readiness for ML. Radiant Earth is actively working to develop Earth observation machine learning libraries and models through an open source hub that support global missions like agriculture, conservation, and climate change. Radiant Earth also fosters a community of practice to develop standards around machine learning for Earth observation and provide information on the progress and innovation in the Earth observation marketplace.

Several discussions about Benchmarks and Training Data sets are captured in these quotes:

- "The quality and source of training data really is a key issue. Also identifying what the correct the best or the good sources of data both data sets and data sources really are and we thought it was interesting that the level of confidence in the data and the outcome is related to the application some applications need more and some less level of confidence.""
- "Building on a geospatial image repository is not as simple as ImageNet."
- "HERE technologies talked about the challenge of maintaining a training set of data that has a temporal characteristic. They need to continuously re-annotate and continuously look to make sure that you got a representation of ground truth."
- "I can't tell you how often we've built a model based on synthetic data with exciting results and then we threw real data we're very disappointed"
- "A theme of our discussion was sharing more whether it's data or whether it's modeling but then also making sure that we have an idea of what the quality is and you know how stale is our data and how good is your model and being able to communicate that as well."

7.4. Augmenting Machine Learning

The previous sections have discussed the hype and challenges associated with AI and Machine Learning. The Training Sets and Benchmarks describes concrete methods underway to improve geospatial machine learning. Additional ideas for addressing the challenges and improving machine learning were also discussed.

- The Role of Domain Experts. Clause 4 discussed the role of domain experts as members of multi-disciplinary data science teams. Domain experts can play a key role in the effectiveness of machine learning. Jay Theodore discussed how we have to solve important problems if we need

to make this trend useful and for that what we need is domain expertise; Without domain expertise we cannot make AI come to life in a meaningful way.

- Humans in the loops. Nils Lahr described how they built algorithms with humans plugging in their expertise at different levels in the overall ML process. The humans provide input that ML can't do. With his example of basketball analysis, there was a real-time/court side loop, along with the upper cognitive layers that come about five minutes after the game concludes.
- Finding a Balance. Regan Smyth described how there is a balance between tweaking the computing to optimize your outputs and tapping into that human knowledge that's a little bit more variable. That's something we're thinking about a lot is can you use the input you receive as initial iterations to figure out systematically what's going wrong and update your methods to address it or do you somehow need to structure your data.
- Combing the parts. Use a combination of use machine learning for specific parts of the model not for the whole pipeline. You'll be able to automate retraining specific parts of the model but not the entire model. No model should go without being paired with a calibrated eyeball. Too often people read in the machine learning as human replacement when in reality it's a force multiplier.
- Theory Guided ML. Yolanda Gil presented work by [Kumar et al 2017] and [Karpatane et al 2017] Kumar and his group at Minnesota incorporating knowledge about physics that constrains what the machine learning models learned. The application was to land use and what kind of crops grow in different areas. They've created this concept of virtual gauges for the river that includes knowledge about physical constraints. Karpatane considered Physics-Guided Neural Network where the learning is consistent physically with what's going on in physics. In this way, knowledge about the world guides the machine learning methods to do better. More will be discussed on this topic in Clause 8.

7.5. Look forward to the future of AI

Yolanda Gil recommended that we consider A 20-Year Community Roadmap for Artificial Intelligence Research in the US. Decades of research in AI have produced formidable technologies that are providing immense benefit to industry, government, and society. AI systems can now translate across multiple languages, identify objects in images and video, streamline manufacturing processes, and control cars. The deployment of AI systems has not only created a trillion-dollar industry that is projected to quadruple in three years, but has also exposed the need to make AI systems fair, explainable, trustworthy, and secure. Future AI systems will rightfully be expected to reason effectively about the world in which they (and people) operate, handling complex tasks and responsibilities effectively and ethically, engaging in meaningful communication, and improving their awareness through experience.

In the near term we conclude this clause with the Keys to Success for Machine Learning presented by Regan Smyth, based on her NatureServe projects:

- Standardized, ground-truthed Training Data
- Partnerships between tech and front line actors
- Human-mediated review of ML outputs

7.6. Recommendations

It is recommended that the OGC GeoAI Domain Working Group consider:

- Promoting development of OGC Community Practices for geospatial machine learning.
- Promoting development of training sets and benchmarks for geospatial machine learning, e.g., in coordination with ESA and Radiant Earth Foundation.

Chapter 8. Tools: Models and Decisions

This clause addresses how knowledge-powered data science can improve decision making.

This Clause contains these sections:

- Knowledge Powered Data Science
- Integrated Modeling
- Model Integration Framework
- Model-supported Decision Making
- Recommendations

8.1. Knowledge Powered Data Science

Yolanda Gil, USC, presented on "Knowledge-Powered Data Science for Integrated Modeling in Geosciences." The call to action for this presentation was encapsulated in this outline:

- Focus shifted from data to models
  - Model Characterization, reuse, and integration
- Need to incorporate model-centered science knowledge about phenomena and context
  - Knowledge about physical, geological, chemical, biological, ecological, and anthropomorphic factors.
  - Knowledge about the user goals and context
- This would enable novel forms of reasoning, integrating, visualizing, managing, learning, and discovery with geosciences data.

Yolanda discussed how this is a very big shift. We need to address the challenges in characterizing reusing and integrating models. To harness data and to develop useful models for decision-making we need to incorporate a lot of science knowledge about these models and the data that they use. We need a diverse set of models for accessible knowledge about physical, geological, chemical, biological, ecological, social, economic factors. Its knowledge regarding what a user might want to do with that model. By infusing our systems with more information and more knowledge using these models will enable new forms of reasoning, integration, visualization, management, learning, and discovery, about the geosciences data at large.

Annie Burgess, ESIP Federation summarized the modeling discussions as focusing on adding knowledge and how it enables more automated and less manual modeling. What works is the idea of provenance and capturing that provenance in models. Everything within this modeling needs to have a URI to point to an explanation of the models.

8.2. Integrated Modeling

There always has been great demand but there’s always been this human bottleneck of manually setting up and integrating these models it takes months or years to put together models for these

kinds of problems and it's really a craft very few people understand how to do this and so it's very hard to integrate a pair of two natural models.

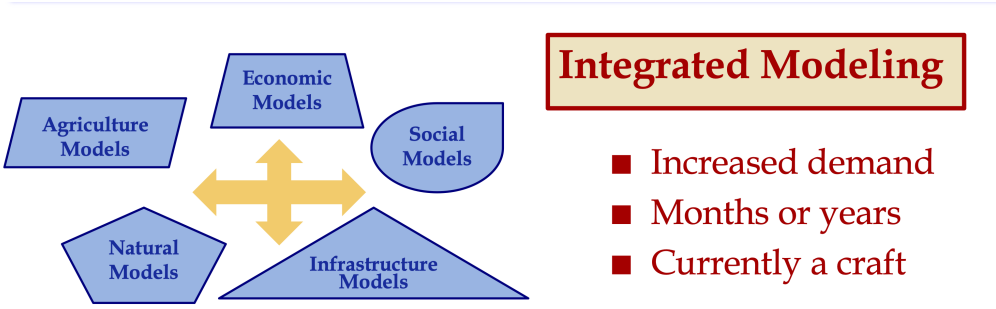


Figure 26. Grand Challenge for Geospatial Data Science

Its really, really, hard to integrate models across disciplines even if both are physical models. Imagine if you try to integrate social economic models with agriculture models. I We'll have a better world if we manage to do integrated modeling. If we can do integrated modeling well then we can do many other things really well. We can do data preparation really well; we can do single model reuse really well; etcetera. This is a grand challenge; worth doing.

8.3. Model Diversity and Integration Challenges

Diversity of models from different disciplines is required to meet the data science and decision needs. With the thematic diversity also comes a diversity of modeling approaches. Integration of the models must address these different types of diversity

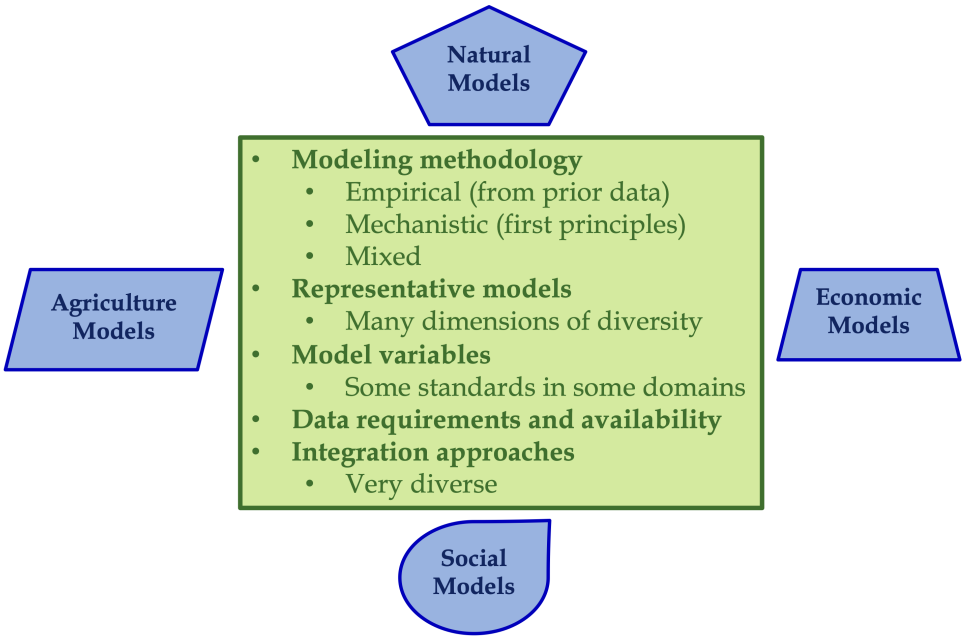


Figure 27. Modeling in Different Disciplines: Diversity of Approaches

Hydrology models may be using an irregular grid where they can look at physical variables in each one of those grid cells. There's a lot of physics involved and the complexity can be enormous. But the system can also be simplified and modeled on a coarser grain level. It needs a lot of historical data to adapt the general physics and fluid equations to that particular area. This generates very useful information about the times and days where particular grid cells are flooded so they're very

important but they have this flavor of very rich physics based modeling.

In contrast agriculture models tend to be more focused on biophysical processes. The growth of the plants, weeding practices, different crops behave differently. These models look at different versions of the crops or different genetic variants and bio geophysical processes. So this is not so much the physics but their processes that are dynamic.

Social models tend to look at societal behaviors through agent-based modeling where you have different groups of agents doing certain behaviors. You can define groups of agents that have children and the children will go to school and so they're able to do the farming or something else. You define all of these behaviors and you see the dynamics of how the system evolves and behaves over time.

So if you're trying to understand and integrate two of these models they work as such different scales they have such different methodology. Some of them are based on theory, some of them are empirical, some are modeling variables that are very different in the physical world. For some with more data and more types of data, they do a better job. But there's not so much data availability and the the ways in which you integrate two models from that both look at physics is very different from the way that you would integrate with a social model. So the challenges are many.

8.4. Model Integration Framework

Research has been done to develop a framework for integrated modeling. We need the ability to incorporate knowledge into our data science systems to improve the way that we do modeling.

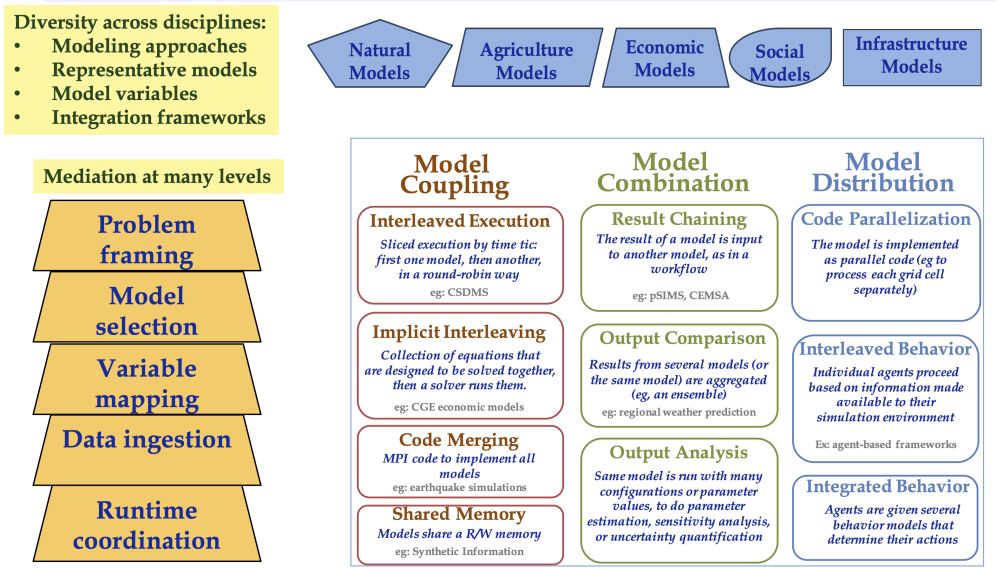


Figure 28. Integrated Modeling: Bridging Across Disciplines

The framework includes several levels in order to address the diversity. The framework was developed in the MINT project which provides Model Integration through Knowledge-Rich Data and Process Composition.

8.5. Model-supported Decision Making

The major value of model development is to improve decision making. The models represent the

accumulated knowledge that can them be applied to the decision making process. The models become a key element in the iterative process to model, analyze, judge, chose or repeat.

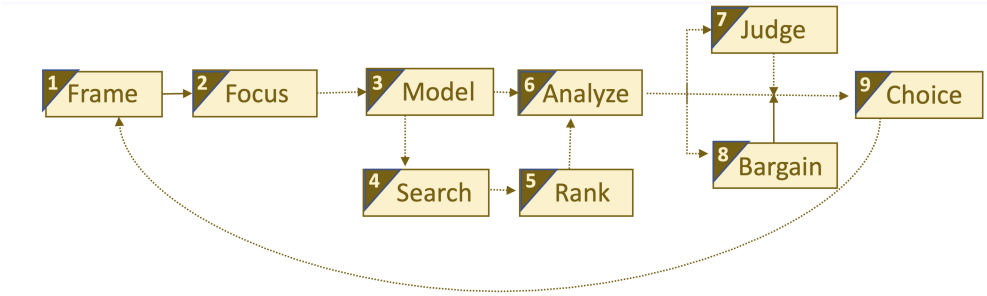


Figure 29. Decision Making with Models

Yolanda's research involves the addition of a key element in decision making - interventions. As as part of the iterative decision making loop, different values of input parameters can be introduced as intervention towards affecting outcomes. Varying values in subsequent model runs provides a range of alternative outcomes. The decision making this informed by the choice against alternative outcomes.

For example in a crop forecasting model, the adjustable variables may be fertilizer costs. By studying a range of fertilizer costs, avaiable budget and the resulting crop production, a decision is informed by the knowledge driven framing based on model indicators and adjustable variables.

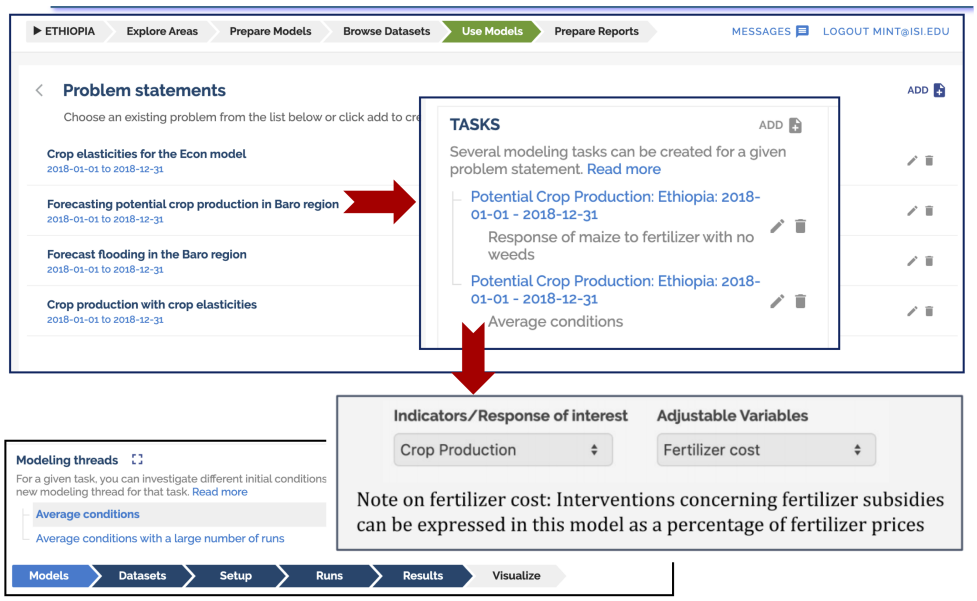


Figure 30. Knowledge Driven decisions

8.6. Recommendations

- Identify the needs for consensus standards in the MINT model integration framework, e.g., data formats coming from different disciplines.
- Expand the discussion on Knowledge Powered Data Science to additional types of models, e.g., models for the built environment, models for training, simulation and gaming.
- Engage the OGC Interoperable Simulation and Gaming DWG in discussion of Knowledge Powered Data Science

- Update the Model, Simulation and Prediction Roadmap in the OGC Tech Trends based on this Clause.

Geospatial Data Science recommendations for the OGC ISG DWG

- Promote development of Knowledge Powered Data Science to additional types of models, e.g., models for the built environment, models for training, simulation and gaming.

Chapter 9. Data Science Applications and Ethics

This clause provides examples of the application of geospatial data science. The objective is to show the value of the data and tools discussed in earlier sections. With the application of the data science technology, ethical issues emerge.

This Clause contains these sections:

- Natural Resources and Agriculture
- Urban Data Science
- Emergency Management / Disaster Response
- Health
- Intelligence
- Business and Insurance
- Geospatial Data Science Ethics
- Recommendations

9.1. Natural Resources and Agriculture

Patrick Giffiths, ESA, provide an example of agriculture monitoring using data science. The motivation for the Copernicus program to support European policies. One of the major European policy elements that has benefited from the Sentinel data is the Common Agricultural Policy. CAP is the single biggest fiscal budget item in Europe at 43% of the total budget of the European Union. Mainly it pays subsidies to make agriculture economically viable for the farmers. But it also has elements for enforcing environmental policies and sustainable agriculture. The CAP compliance checks for receiving subsidies is traditionally done through interpretation of a very high resolution imagery for a 5% sample per country and then for a smaller percentage of those parcels there is spot checks in in addition to the VHR imagery interpretation to determine the compliance. This is complicated and inefficient. Fully automated monitoring based on Sentinel time series and data analytic tools basically from the data science community is a big step. The Figure from ESA's Sen4CAP project shows the results of this new process.

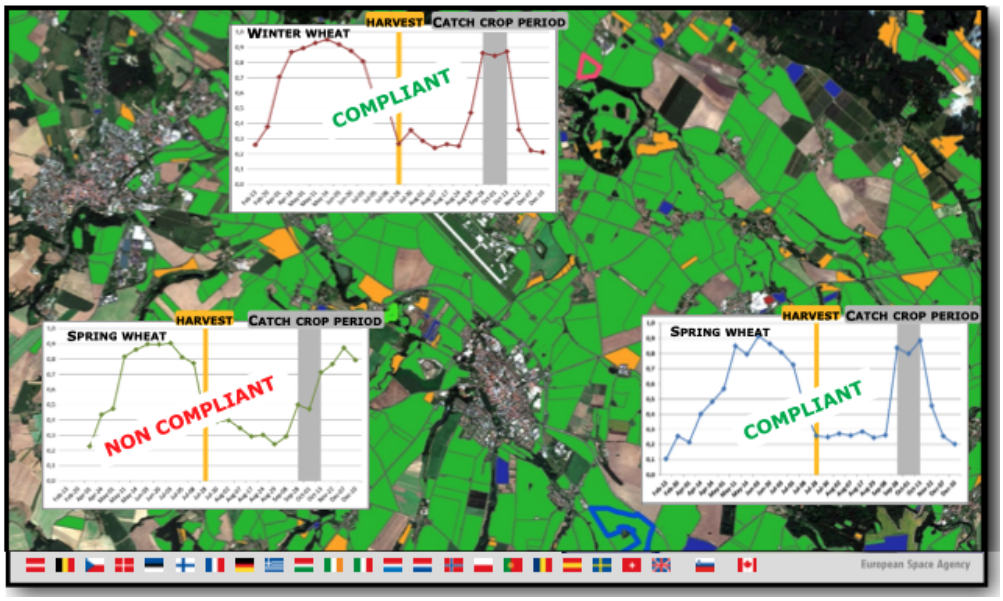


Figure 31. Machine Learning supported Agriculture policy compliance

Recall these example applications presented earlier in this document:

- Crop inundation application in the Geographical Statistics Clause 6.
- Biodiversity application in the Machine Learning Clause 7

9.2. Urban Data Science

The City of Los Angeles has several data science projects improving the lives of urban residents, as described by Jeanne Holm, City for Los Angeles:

- SmartAirLA/Predicting What We Breathe collects data from satellites airborne instruments and ground data sensors from Internet of Things across the city of Los Angeles and across the region. The data is federated in a way that helps to use machine learning to create predictive analytics about the quality of air. LA is partnering with other mega cities around the planet to see patterns can be evoked from satellite data to understand things that are happening on the ground. In Mumbai for example where they may not have as many ground data sensors, it's a pretty interesting experiment.
- ShakeAlert LA is an earthquake early warning system. It gives you up to minutes warning that you are about to feel shaking from an earthquake. In Los Angeles and so we have over a million users and they partner with the USGS ShakeAlert system which has sensors up and down the west coast. The app technology and latency and the network was tricky, but it's a lot about behavioral nudges and behavioral science to get people to be safer during an earthquake.

NVIDIA Metropolis create IoT applications, from the edge to the cloud, for retail, inventory management, traffic engineering in smart cities. Milind Naphade, NVIDIA Metropolis, presented at the LP_DS Summit about their system for workloads that run at the edge for video data. Processing video for traffic monitoring and analytics detects objects at 50 objects per frame that's 1.5 million messages per second. With video plus geospatial every object that detected is labeled with GPS coordinates, we convert the video into the real world map based understanding. This for example allows for anomaly detection, e.g., where a vehicle was detected driving in the wrong lane. This would allow for real time response like switching the traffic lights to red, police vehicle escorting

the vehicle off of the road, etc.

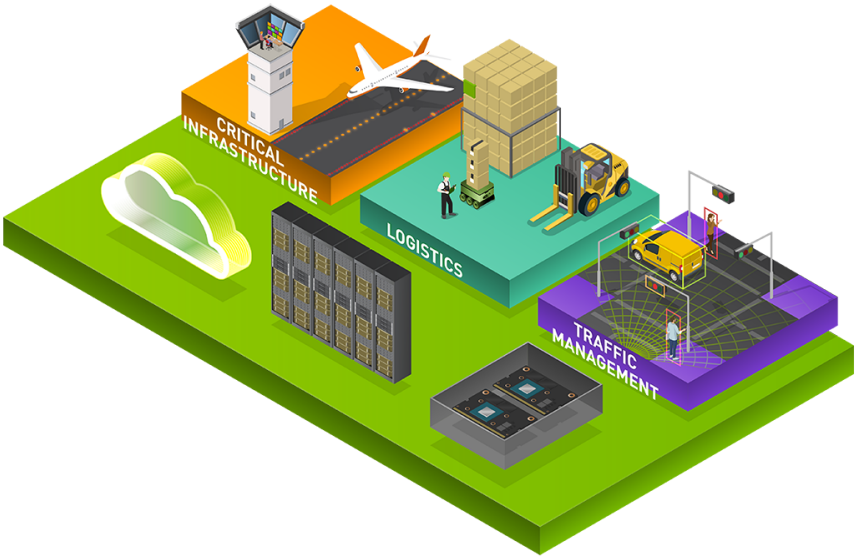


Figure 32. NVIDIA Smart Cities

Ed Strocko, US DoT, provided a focus on transportation safety. Over 36,000 people die on highways every year. This is way too high. The five examples in the Figure are from a data visualization organized by DoT. The Central Florida example used machine learning to do short-term predictive analytics on where a crash is going to occur; where do we think we need to be focusing our attention changing the variable speed limit signs; giving the infrastructure operators some information. Ford and Arity were looking at the behavioral information coming off the cars; hard stops, hard starts, people using cell phones and using some data science in there to get that get down that number of fatalities. These developments need to continue until we can really achieve that that vision of a much safer road with autonomous and connected vehicles

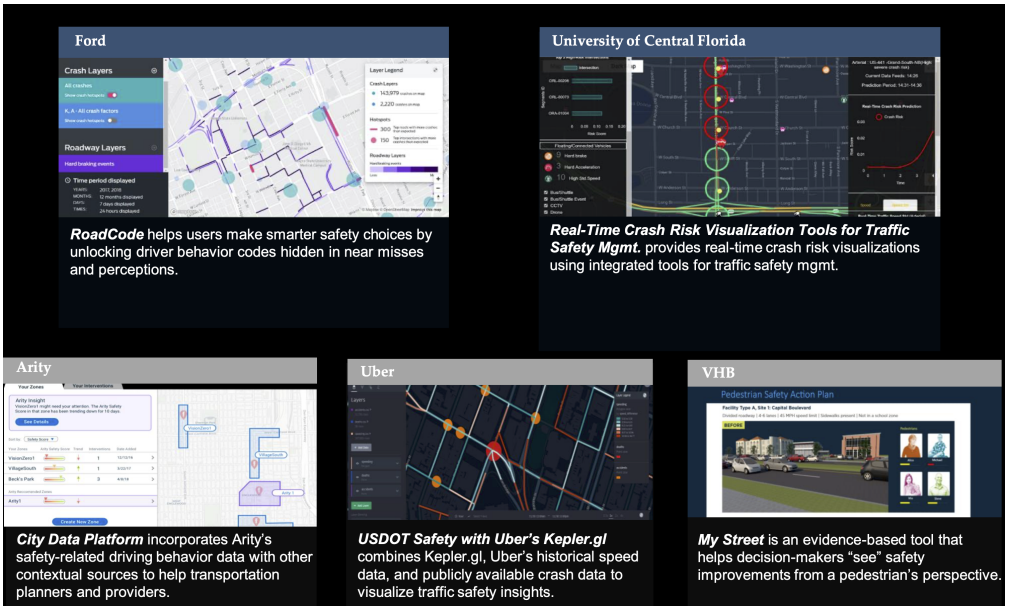


Figure 33. US DoT Transportation Safety Data Visualization Challenge

To support many of the urban use cases, High Definition (HD) Maps are needed. Standardizing HD maps is contrasted with today's mapping systems. Information for autonomous vehicle needs to be HD. Standard for HD map data in a format that can be understood by not just the cities but software in general and the other scientists at large. We need to look at the policy comes together. But it's definitely the collaboration across the spaces there to get to something. Jeremy Morley, Ordannce

Survey UK, discussed how from a national mapping agency perspective as well for the UK, there's interesting questions also as to whether when we talk about HD roads or just interested in that CAV market. It's not just a single map that each of fleets will want. And how well does it serve other purposes as well whether it's the IOT market or simply local authorities better maintaining their assets. There's a way to go to accumulate enough evidence as to what is a good product or standard in this space.

9.3. Health

Stephanie Shipp, U. of Virginia, presented on Harnessing the Power of Data to Support Community Health and Well-Being. She described the Community Scapes program that identifies where to target programs and policies for communities with risk of obesity. A key part of the data analysis was re-distribution of source data using synthetic information. The project used American Community Survey (ACS) summaries and ACS Public Use Micro Data Sample (PUMS) to impute synthetic person data for all people or households in area of interest. The data was re-weighted synthetic data according to ACS tables to simultaneously match the relevant distributions, to Census Tracts or Block Groups. The aggregate synthetic data was used to compute summaries, and margins of error, over the new geographic boundaries or interest as shown in the figure.

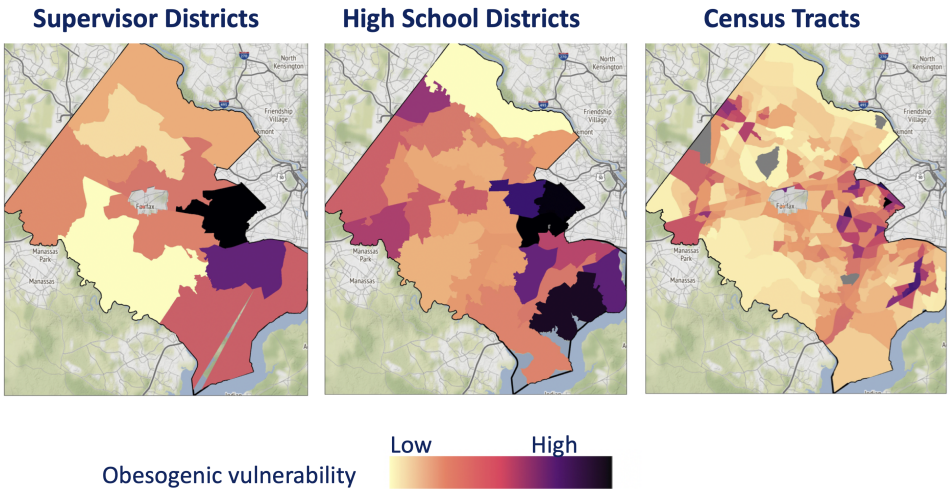


Figure 34. Identifying communities with risk of obesity

Wendy Martinez, US BLS, described how the US Center for Disease Control maintains an environmental Public Health tracking network with information on environmental hazards and the health effects associated with the hazards. Such health data can be applied for planning and health interventions.

Ajay Gupta, led a group discussion on health that examined the need for both population level and precision precision level location data:

- utilizing patient-level individual data to understand their daily exposure to environment what type of environment whether it's physical environment or nature environment including air quality noise and how does that impact their health outcome.
- Predictions using GeoAI technology to understand the moment of eating. With heart failure patients it's very important to identify high propensity of eating in order to suggest nearby healthier options.

- The Supercomputing Center at UC San Diego is utilizing satellite imageries and sensor survey data to understand a neighborhood characteristic and the risk factors that directly impact certain health or certain disease outbreaks.

9.4. Intelligence

Nils Lahr provided an example from the National Geospatial-intelligence Agency (NGA). The NGA workflow for real-time video from UAVs is shown in the Figure. The system manages processing of a hundred drones. The analysis is about where and when. Gathering patterns of life not just as whole bunch of data points but things that matter in the field. The patterns are a level of intelligence that we have not been able to do at scale.

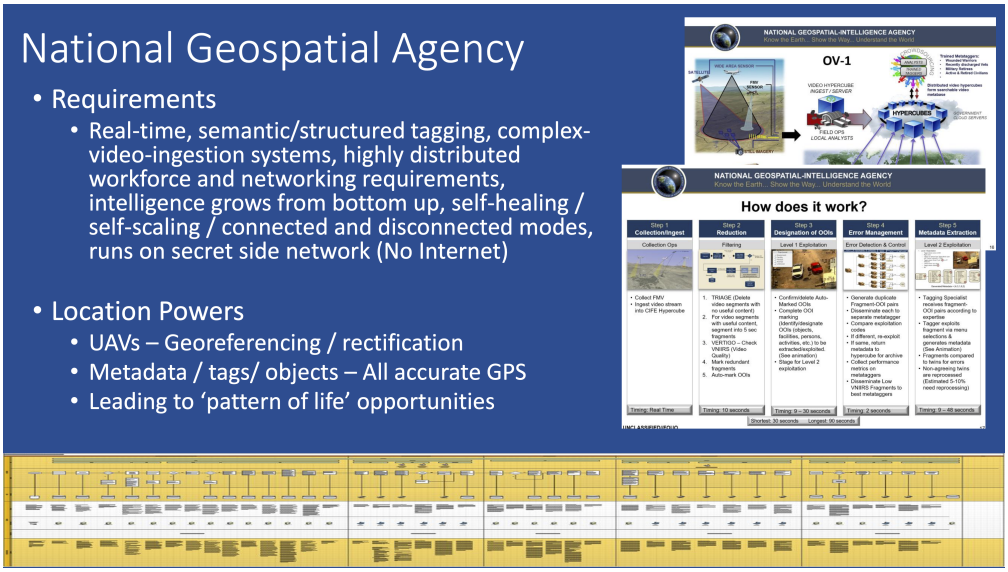


Figure 35. NGA drone video processing workflow

9.5. Business and Insurance

Nils Lahr presented several examples from the world of business

- Hedge funds are making use of video analytics of Walmart parking lots. Public data feeds on Black Friday before Christmas, or maybe even three months before, to start predicting how people are going to spend money at certain times. Algorithms are used to figure out to place bets on the future. In this case it's not only where the cars are located in terms of the globe but literally where they are in the parking lot.
- Logistics in handling new automobile shipments involves moving ten thousand cars per day on thirty acres of parking lot. Need to plan for what cars are leaving tomorrow to get them close to the train. This is a localized geospatial application: how do you tell the one guy to go to the one car that he needs and how do you know the car is there. The input video comes from UAVs reading the VIN numbers along with light posts with wide area cameras. Geospatial information essentially GPS is used to create a digital twin of the parking lot. The digital twin is used to optimize the logistics of moving the cars.

An LP_DS discussion group focused on geospatial data science applied to insurance:

- Satellite imagery is helping the insurance industry you know a lot of for example the use case is

underwriting

- Insurance policy applications are includes much risk information that is pre-populated based on address.
- Use of geospatial information for the allocation of claims adjusters along with a rough sense of the damage before going to the site.
- Catastrophe modeling, e.g., earthquakes, flooding, in advance for risk assessment. Requires a very high level of accuracy, e.g, for just a few meters for the water damage
- Insurance is very location-based. With all the information, from imagery and from modeling, enriches the underwriting process.

9.6. Emergency Management / Disaster Response

Jay Theodore, Esri, asked the question of an LP_DS panel: "what's the most meaningful and satisfying project you've been involved in applying data science?" Devaki Raj responded about the application of data science to disaster response.

Devaki Raj, CrowdAI, provided examples of applying machine learning to the to hurricanes in Houston and Florida; and about the Santa Rosa and Campfire Fires in California. She spoke responding to the largest operational challenges that often occur after major natural disasters. CrowdAI uses different types of third-party imagery, e.g., satellite, drone, aerial.

- Hurricane assessments. With Hurricane Harvey CrowdAI mapped all the roads on imagery prior to the event; and then mapped all the roads on post flooded imagery. This roads condition were converted from TIFF file format into GeoJSON vectors and provided to first responders. For Hurricane Michael in Florida, NOAA aerial imagery was used to identify building damage based on mapping of almost 18,000 buildings in a couple of minutes.
- Wildfire assessments. For the Campfire wildfire in California they applied a model that had been trained with Digital Globe imagery from 125 countries. After Campfire, 25,000 buildings were mapped off of the imagery that Digital Globe had on their open data platform. At 30 centimeter resolution, pre-fire buildings were mapped as 25,000 polygons. The post fire assessments were at the level of individual houses as the aggregate was not useful. In the same neighborhood some buildings were standing and some are completely destroyed. Mapping after the Santa Rosa fires was used for risk mitigation. Imagery analysis with machine learning identify wildfire risk factors for a future fire. This was to mitigate risk potentially from a future disaster.

The application of Statistical Geography provides data science methods to assess the social impact of a hurricanes. The Bureau of Labor Statistics mapped the affect of storm surge due to a hurricane hitting Virginia coastal areas. Using the Quarterly Census of Employment and Wages (QCEW), they calculated the employment in the various geographical flood zones.

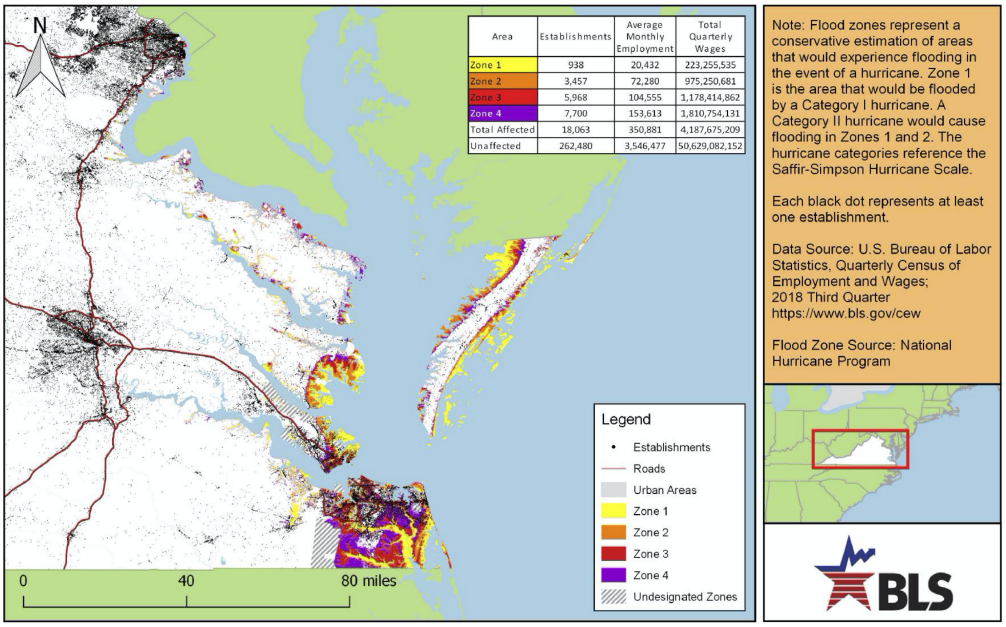


Figure 36. Employment in Hurricane Storm Surge Flood Zones, Virginia

Examples of emergeing response from previous clauses:

- Near Real-Time Flood Mapping of Agriculture by the USDA National Agriculture Statistics Service as in the Statistical Geography section of Clause 6.
- Mapping Flood Inundation at continental scale in the CyberGIS section of Clause 6.

9.7. Data Science Ethics

With the application of data science comes the question of ethical use of the data and the associated analytics. The technology and data we have described in earlier paragraphs could be applied in a variety of ways good or bad. It's with the application of technology that issues of ethics arise.

```
Technology is neither good, bad; nor is it neutral
- 1st Law of Technology (M. Kranzberg)
```

Wendy Martinez, provided an outline of Data Ethics. She described Ethics as: the study of right and wrong; as the set of moral principles governing our behavior; and as often abstract, guidelines. Data Ethics is a "branch of ethics…moral problems related to data, …algorithms, … and corresponding processes.

Three Axes of Data Ethics:

- Ethics of Data: Collection and analysis of large datasets
  - Re-identification of individuals - geospatial concern?
  - Trust and transparency
- Ethics of Algorithms: Increasing complexity and autonomy of algorithms (e.g., Internet of Things)
- Ethics of Practices: Responsible innovation, R&D, usage - foster innovation and protect rights

- Informed consent (Web-scraping??)
- User privacy and surveillance
- Secondary use - integration of data sets
- Unintended use

Wendy provided three examples were discussed. The first was on racial bias in medical algorithms. The algorithm underestimated health needs of sickest black patients. Mapping highest scores showed concentration in affluent suburbs. The second was on Predictive policing software. The software focused on already hotspot areas, leading to geographic profiling. Adding police resulted in an increase in reports. The resulting spike was used as justification. The third was on autonomous vehicles, what should a vehicle algorithm do when faced with several undesirable choices.

References on data ethics.

- Self-driving car dilemmas reveal that moral choices are not universal
- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
- Code of Silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out.
- Racial bias in a medical algorithm favors white patients over sicker black patients
- A face-scanning algorithm increasingly decides whether you deserve the job

Data Science Ethics is not just an abstract discussion. The application of data science technology can bring harm. The questions about technology can prevent its application. Philippe Cases, Topio Networks, discussed ethics and principles in the context of AI and edge computing: we cannot compare the new technology to some absolute standard; the appropriate approach is to identify the advantages and minimize the risks.

Andy Brooks discussed the intersection of AI and ethics. With regard to counterterrorism its a discussion about targeting and lethality. It's not just an academic discussion or model. What's the ethics of using automata to do a certain type of work?

Andy Brooks discussed the implications of AI on workforce. Previously it would take say ten people two weeks to do one thing, and now it takes one person clicking on a script and its done in ten minutes. There's a lot of implications for that with regard to employment and workforce and staffing.

Ethical issues particular to geospatial data science are highlighted by handling of location data and location privacy. LP_DS discussed that it is very difficult to make data anonymous when it contains location information about individuals. In particular trajectory data about individuals has been shown to be de-anonymized rather easily. Strategies for data ethics were discussed such as relating to edge computing. Keeping the most descriptive data on individuals at the edge and passing the most general information to the cloud was discussed. There was also a call for a Geospatial Data Science Code of Ethics. Some of the approaches suggest masking or otherwise degrading the data. Stephanie Shipp advocated that when it comes to data privacy, don't mask the data but rather punish misuse.

9.8. Recommendations

- Identify and promote additional applications of geospatial data science.
- Identify and promote community practices for geospatial data science ethics.

It is recommended that the OGC GeoAI Domain Working Group consider:

- Promoting development of a Geospatial Data Science Code of Ethics focused on Artificial Intelligence.

Chapter 10. Emerging Trends

This Clause addresses these topics:

- Edge Computing
- Heterogeneous Computing
- Technology Forecasting
- Recommendations

10.1. Edge Computing

The Internet of Things is anticipated to connect billions of devices to the internet. We are still in the early stage of IOT deployment, this according to Philippe Cases presentation at LP_DS. IOT adoption was expected in 2016/2017. Data now suggest otherwise as many companies have yet to roll out their products. Perhaps what was missing was computing at the edge which now seems emergent. During the Edge Computing World conference in December 2019, Amazon Wavelength was announced as bringing AWS services to the edge of the 5G network. Jay Theodore, Esri, anticipates several types of Edge Computing: Edge Servers, Edge Portals, Edge Devices.

Tradeoffs in distributed computing will be needed when considering computing at the edge and cloud computing. In Clause 3, we saw how data at the edge is big data with estimates, e.g., 4TB/day for each car. When you're sending massive amount of data through the network, connectivity becomes an issue. Cloud computing is very effective for some applications as an on demand elastic heterogeneous measured service. These have been very effective for geospatial computing at scale. Now computing at the edge is becoming more powerful, for example, NVIDA is developing a processor which can be deployed to the edge and execute at 21 trillion operations per second. But there's a real problem with cloud computing and that is latency. Communication even at the speed of light is too slow over distances for certain applications. So there's a compromise as described by Marc Armstrong, of edge and fog computing to put processing in between IoT devices and the cloud in order to cut down on the amount of data that's transformed or transferred. Jay Theodore suggested that for batch inferencing you probably go to the cloud, and if you want to do interactive inferencing you perform that at the edge. It turns out that geography really does matter when considering where to perform distributed computing.

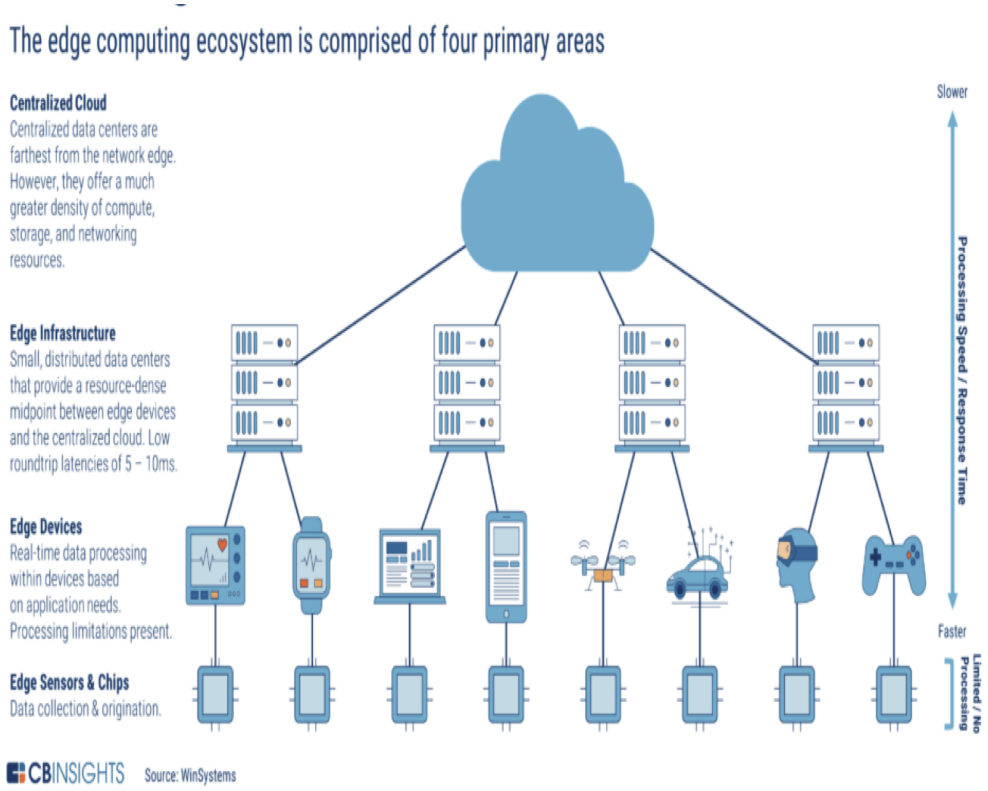


Figure 37. From Edge Sensors to Centralized Cloud

The Edge is also important because its location and presence over time. The sensing and actuation associated with IoT devices at the edge can make use of the location. The data obtained from sensing at the edge has a location content that may be useful immediately in edge computing and the the location can be attached to the data as it is distributed to the cloud and other devices. Philippe Cases's surveys indicates that much of the sensing at the edge is video and that of all the data at the edge, 75% of the data is actually time series.

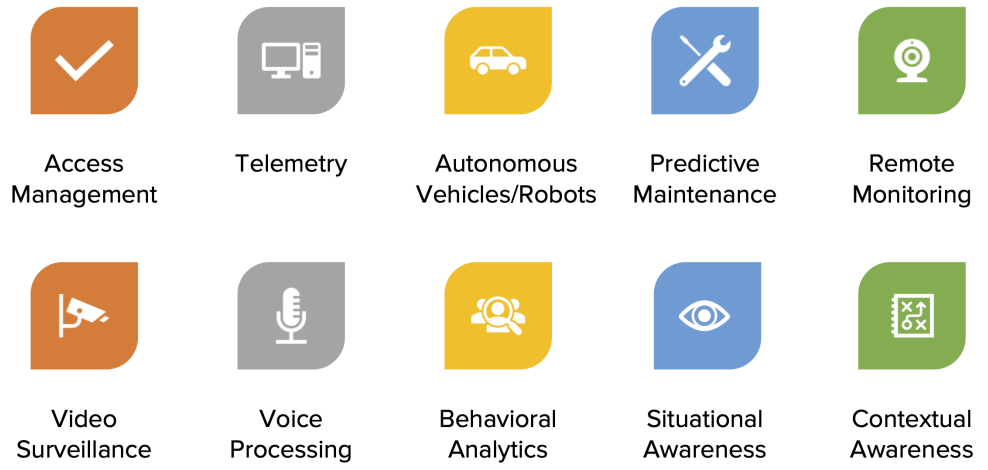


Figure 38. Key use cases for processing at the edge

Marc Armstrong discussed data from the edge as "Fast Data." Fast Data streams are driving computational complexity that needs to be addressed by experimental architectures and new analytic methods. The roots of fast data come from research to temporal and dynamic GIS (See Clause 6). We now have fast streaming data that is orders of magnitude faster. Experimental architectures to develop new methods are needed to do analysis of large and fast data and time

frames. Challenges of fast streaming data include, unknown sample size, non-stationarity, and algorithmic complexity in space and time. Techniques such as reservoir sampling and approximate computing are suggested by Marc Armstrong for developing fast data analytics.

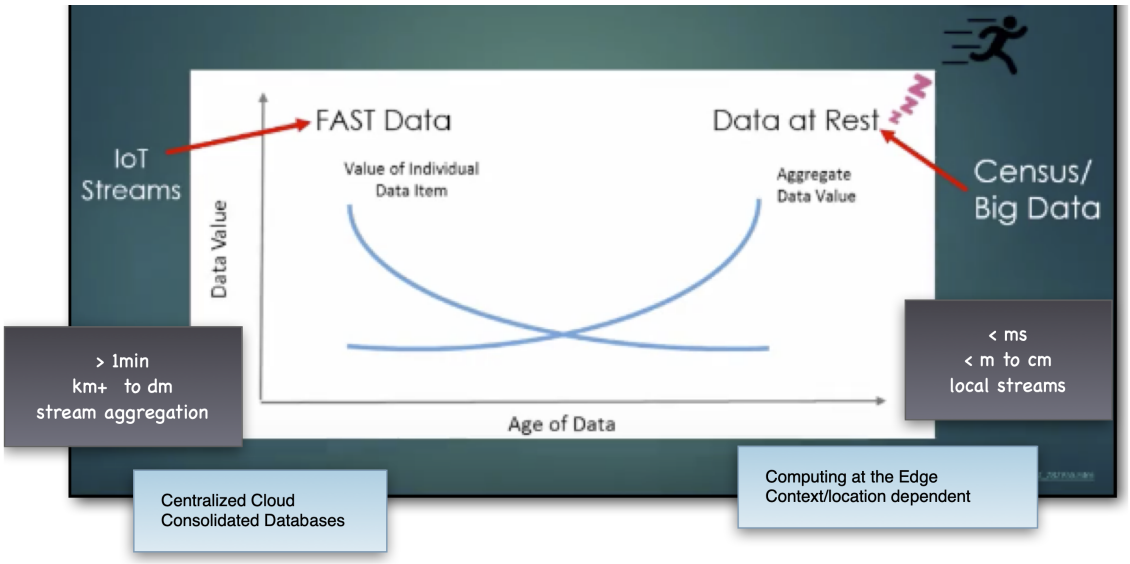


Figure 39. IoT sensors Create Fast Data Streams

Coordinated observations is a powerful use case based on the emerging architecture of IoT sensing and actuation, Edge Computing, Fast Streaming Data, 5G communications and cloud computing. Observations by an IoT sensor used to trigger subsequent computing, additional observations and subsequent actions all done in the distributed network without human observation. This observation-processing-actuation workflow in the network suggests a powerful reusable pattern:

- Feature of interest detected in video in an initial location.
- Feature fits a decision rule that triggers subsequent observations or processing associated with the initial location or in other locations based on the trajectory of the feature of interest.
- Actuation of some IoT device that changes the processes associated with the detected feature of interest, again, either in the initial or other location.

This pattern can be extended based on the decision fusion pattern discussed in the OGC Fusion Study: useable templates of "If-This-Pattern-Consider-This-Decision."

This discussion on Edge Computing at LP_DS can be used to update the OGC Tech Trend on Edge Computing.

10.2. Heterogenous Computing

Marc Armstrong presented about the opportunities of Heterogenous Computing to geospatial data science. Heterogeneous computing is a scheme in which the different computing nodes have different capabilities and/or different ways of executing instructions. In heterogeneous computing, the cores are different. The figure shows a heterogeneous system with multi-core, GPU, FPGA, etc. We have seen use of GPUs earlier (Clause 6) for accelerating geospatial analytics. Tensor Processing units are an another computing architecture that has value to geospatial.

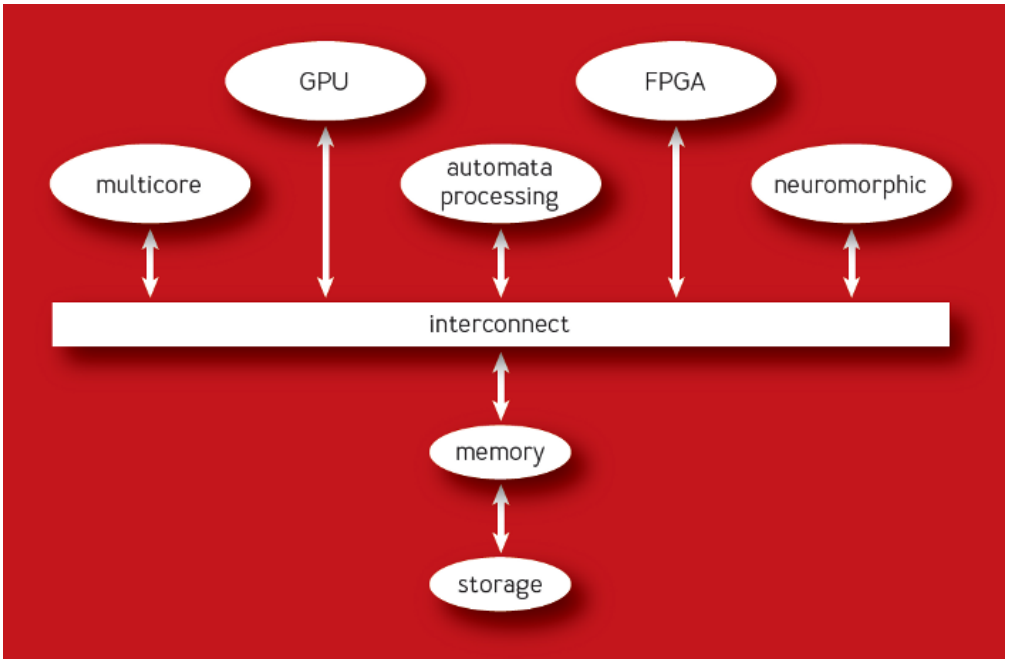


Figure 40. Generic Heterogeneous System

Some parts of geospatial problems are addressed with different computing architectures. We can anticipate developing spatial middleware that would align the characteristics of geospatial algorithms to particular types of Hardware environments. OGC is already working with the Khronos Group for geospatial computing based on GPUs. Working with the Heterogeneous System Architecture (HSA) Foundation in a similar fashion could bring additional improvements in geospatial data analytics.

Heterogenous Computing will be added to the OGC Technology Forecast based on the discussions at LP_DS.

10.3. Technology Forecasting

OGC conducts a forecasting activity for geospatial technology. The forecasts provide early identification of disruptive technologies; supports discovery-driven planning; and drives OGC member decisions regarding geospatial innovations. The figure shows the a summary of the process.

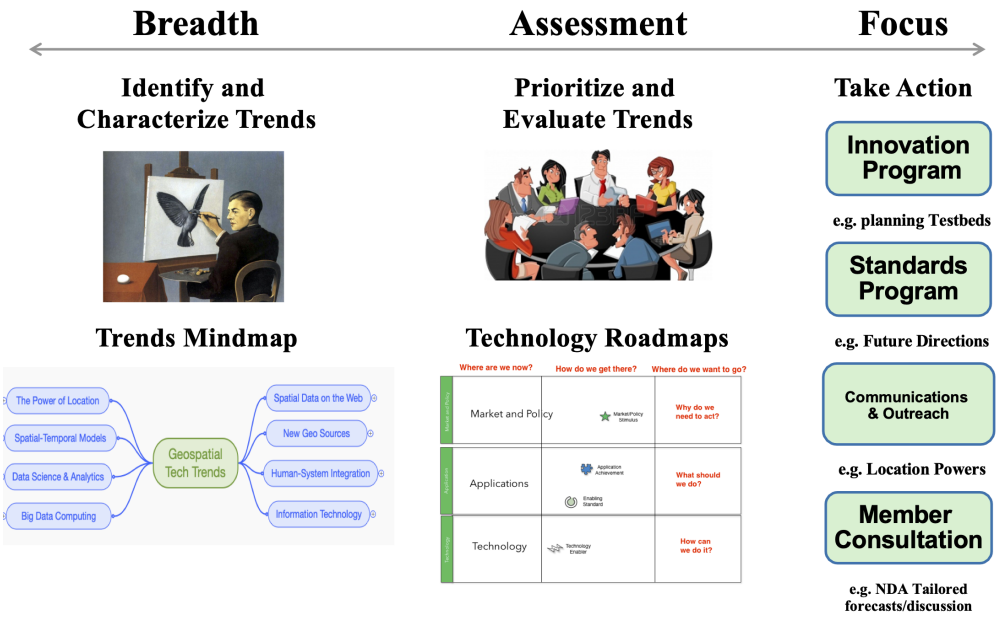


Figure 41. OGC Technology Forecasting

The concept of the Location Powers: Data Science Summit and for some of the other Location Powers summits came from analysis in the OGC Technology Forecasting program. The Forecast will be updated based on the results of LP_DS and a will drive discussion about the focus of future Location Powers events.

10.4. Recommendations

- Propose use cases for Edge Computing in the OGC Testbed planning.
- Discuss applications and computing methods for fast streaming data in OGC working groups
- Add Heterogeneous computing to the OGC Technology Forecast
- Consider heterogeneous computing as a topic for Future Directions Session.
- Review the work of the HSA Foundation for possible discussion topics on geospatial computing.
- Develop concepts for future Location Powers: Summits based on Location Powers: Data Science results.

Recommendations regarding Edge Computing are suggested to the Moving Features SWG and Temporal DWG:

- Promote development of OGC Community Practices for spatial-temporal analytics.
- Propose use cases for Edge Computing: temporal analysis of streaming data

Chapter 11. OGC activities on Geospatial Data Science

An objective of the Location Powers series is to identify OGC activities based on the outcomes and recommendations of the summit. In addition to the recommendations listed in previous clasues, this clause identifies potential OGC activities in three areas:

- OGC Program Activities
- External Coordination on Standards
- External coordination on R&D

11.1. OGC Program Activities

- Communicate the results of LP_DS Summit
  - Publish this White paper
  - Present the results of LP_DS to multiple venues
- OGC Standard Program
  - SP Working Groups: Discuss recommendations with WG chairs
  - Future Directions Sessions
- OGC Innovation Program
  - Monitor the related OGC Testbed 16 activities to update the concepts in this white paper
  - Recommend Testbed 17 work items on geospatial data science for consideration by sponsors
- Tech Trends:
  - update mindmap and priority trends
  - Plan for Location Powers 2020 summit based on results of LP_DS

11.2. External Coordination on Standards

- Identify organizations focused on developing standards and best practices in Data Science
- Consider establishing liaisons with external organizations to develop geospatial data science.

11.3. External coordination on R&D

- Introduce Geospatial Data Science as a theme in the annual Apache Software Foundation (ASF) annual conference geospatial track.
- Participate in the NSF Geospatial Software Institute developments as they may happen in 2020 based on the incubator study concluded in 2019.

Annex A: Location Powers: Data Science Summit

This Annex includes these topics about the Location Powers Summit:

- Summary Agenda
- Detailed Agenda
- Organizing Committee
- Participating organizations

A.1. Summary Agenda

Day One - November 13

- Session 1: Foundations
  - Data abundance, big data analytics,
- Session 2: Analytics and Representations
  - Mathematical methods, computer algorithms and data structures
- Session 3: Ripe Trends: AI/ML
  - AI, Machine learning

Day Two - November 14

- Session 4: Outcomes/Applications
  - How can geospatial data science make a difference?
- Session 5: Actions to Take
  - What should organizations do to advance geospatial data science and its outcomes?
- Summary: Rapporteur Reports

A.2. Detailed Agenda

A.2.1. Opening:

- Welcome by Google : Ed Parsons, Google’s Geographer
- Welcome by OGC : Nadine Alameh, OGC CEO
- Workshop Overview: George Percivall, OGC CTO and Workshop Moderator

A.2.2. Session 1: Foundations/Motivations

- Session Overview
  - Moderator: Ed Parsons, Google

    - Rapporteur: George Percivall, OGC
    - Description: With the inundation of new geospatial data sources from imagery, sensors and the potential of new broadband pipelines such as 5G, there is a defined need of validating and cleansing data. These data will be in both structured and unstructured formats. This session on foundations will illustrate the challenges of curating data from these myriad sources and understanding both the veracity and accuracy that will be attributed to these sources.
- Presentation on Foundations
    - Title: Fundamental Issues in Geospatial Data Science: Emerging Trends in Data and Analytics
    - Professor Marc Armstrong, Associate Dean and Professor of Geography, University of Iowa
- Presentation on Foundations
    - Title: Scaling machine learning to handle visual data will result in more powerful AI
    - Nils Lahr, CEO Orion Systems
- Panel on Foundations:
    - Kathleen Stewart, UMCP/CGIS - Perspective: New opportunities through big mobility data analytics
    - Anand Padmanabhan, University of Illinois - Perspective: CyberGIS-Jupyter for Reproducible Geospatial Research and Education
    - Mark Korver, AWS
    - Jayant Sharma, Oracle

A.2.3. Session 2: Analytics and Representations

- Session Overview
    - Moderator: Kumar Navulur, Maxar
    - Rapporteur: Annie Burgess, ESIP Federation
    - Description: While geospatial scientists in the past may have spent 70% of their time on data capture and management, and only 30% on analytics, that model today and in the future will be flipped. As such, the engines of data analytics will need to be placed into the hands of both traditional GIS analysts as well as business intelligence and knowledge workers. This will require new tools to derive spatial statics, as an example, as well as utilizing graph technology for entity resolution. This session will aim to examine these topics and more.
- Presentation on Analytics & Representations
    - Title: Knowledge-Powered Data Science for Integrated Modeling in Geosciences
    - Professor Yolanda Gill, USC Spatial Sciences Institute
- Panel on Analytics:
    - Todd Mostak, OmniSci - Perspective: Geospatial Intelligence through Accelerated Analytics & Data Science at Scale
    - Lauren Bennett, Esri - Perspective: Spatial data science moves from the avant garde to the

mainstream

    - Hamed Alemohammad, Radiant Earth - Perspective: Radiant MLHub: A Repository for Machine Learning Ready Geospatial Training Data
    - Keith Hare, JTC 1 SQL/GQL convener - Perspective: Analytics And Representations in SQL and GQL
- Discussion Group on Analytics

A.2.4. Session 3: Ripe Trends

- Session Overview
    - Moderator: Jay Theodore, CTO, Enterprise Technologies, Esri
    - Rapporteur: K. Kim, AIST and GeoAI DWG chair
    - Description: Computing at the edge will require interoperable sensors and other devices, machine learning and faster transfer of data. The coalescence of technologies such as IoT, 5G and cloud native computing is facilitating a new era of geoprocessing. This will impact the way data is captured, managed and processed where AI portends to impact everything from traffic management to adtech.
- Presentation on Ripe Trends
    - Title: AI at The Edge
    - Philippe Cases, ReadWrite Labs
- Panel on Ripe Trends:
    - Anand Kannan, Pitney Bowes - Perspective: Data science, an interdisciplinary approach
    - Milind Naphade, CTO, Metropolis - NVIDIA - Perspective: AI-IOT and Location
    - Devaki Raj, CrowdAI
    - Jim Stokes, MAXAR

A.2.5. Session 4: Outcomes/Applications

- Session Overview
    - Moderator: Jeremy Morley, Ordnance Survey UK
    - Rapporteur: Ajay Gupta, chair of OGC Health WG
    - Description: The geospatial community recognizes the significance of location-based data but how are these data revealed and recognized as inputs to data science? Is the community supporting the integration of geospatial data with other enterprise computing solutions and ensuring that data scientists understand their value, social effects, applications and expected return on investment.
- Presentation on Outcomes
    - Title: Integration of Geospatial Data: Examples and Implications
    - Dr. Wendy Martinez, President-elect, American Statistical Association; and US Bureau of Labor Statistics.

- Panel on Outcomes:
  - Regan Smyth, NatureServe - Perspective: The Age of Precision Conservation: Applying AI and Collaborative Science to Prevent Species Extinctions
  - Megan Furman, Defense Digital Service, OSD
  - Steven Ward, The Climate Corporation
  - Edward Strocko, USDOT Bureau of Transportation Statistics
- Discussion Groups on Outcomes: trail walk to the Bay
- Reports from Discussion Groups on Outcomes

A.2.6. Session 5: Actions to Take

- Session Overview
  - Moderator: Nadine Alameh, OGC
  - Rapporteur: Adam Martin, Esri
  - Description: This panel will attempt to identify what data science brings beyond traditional GIS and vice versa. What skill sets will be required and how do we train data scientists who are expected to use geospatial data. The objective will be to help organizations automate and scale geospatial data science insights into workflows throughout their organizations. In particular, what actions might OGC take. This final panel will summarize the suggested points of deliberation from previous sessions and suggest next actions.
- Presentation on Actions:
  - Title: When HPC met AI - Next generation of Geospatial Intelligence powered by the ABCI
  - Satoshi Sekiguchi, AIST
- Presentation on Actions:
  - Title: Designing the Future of Data Science
  - Andrew Brooks, NGA
- Panel on Actions:
  - Patrick Griffiths, ESA - Perspective: Earth Observation data and analytics supporting policy and geospatial industries
  - Jeanne Holm, City of Los Angeles - Perspective: Building a Generation of Government Data Scientists
  - Stephanie Shipp, U. of Virginia - Perspective: Harnessing the Power of Data to Support Community Health and Well-Being

A.2.7. Summary Session

Rapporteur Reports :

- Session 1: George Percivall
- Session 2: Annie Burgess

- Session 3: K. Kim
- Session 4: Ajay Gupta
- Session 5: Adam Martin

A.3. Organizing Committee

- Ed Parsons, Google
- Patrick Griffiths, European Space Agency
- Don Sullivan, NASA
- Caroline Bellamy, Ordnance Survey
- Roy Rathbun, NGA
- Kyoung-Sook Kim, AIST
- Tracey Birch, SOFWERX
- Shaowen Wang, University of Illinois - Urbana Champaign
- Kumar Navular, Maxar
- Adam Martin, Esri
- Joe Francica, Pitney Bowes
- George Percivall, OGC

A.4. Participating Organizations

- AAIA
- Aechelon
- AIST
- Amazon
- Arturo
- Ca PUC
- City of Los Angeles
- Cray / HPE
- CrowdAI
- CustomWeather
- DLR
- ESA
- Esri
- Geospatial Alpha
- Haystax

- HERE
- JTC 1 SQL/GQL
- LocusLabs
- Maxar
- NASA
- NatureServe
- NGA
- NVIDIA
- OmniSci
- OS
- PB
- Polaris Wireless
- Radiant Earth
- Stanford
- Topio Labs
- U. Chicago NORC
- UCSB
- Univ of Virginia
- Urban Footprint
- US BLS
- USAF ISR
- USDOT

Annex B: Revision History

Date	Release	Editor	Primary clauses modified	Description
2020-02-28	0.1	G. Percivall	all	initial version
2020-05-20	0.2	G. Percivall	numerous	editorial updates based on participant reviews

Annex C: Bibliography

This bibliography was developed to support planning for the workshop.

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

[2] 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

[3] 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

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

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

[6] 50 years of Data Science, David Donoho, 2015

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

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

[9] “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/

[10] 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

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

[12] 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

[13] 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.

[14] doi: 10.1109/JSTARS.2018.2795753

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

[16] 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

[17] 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