Next-Generation Smart Environments: From System of Systems to Data Ecosystems

Edward Curry
Lero, NUI Galway

Amit Sheth
Kno.e.sis, Wright State University

Digital transformation is driving a new wave of large-scale data-rich smart environments with data on every aspect of our world. The resulting data ecosystems present new challenges and opportunities in the design of intelligent systems and system of systems.

Smart Environments are generating significant quantities of data due to a convergence of digital infrastructure from the Internet of Things (IoT), Edge, Fog, and Cloud Computing that is driving a new wave of data-driven intelligent systems. Through the generation and analysis of data from the smart environment, data-driven systems are transforming our everyday world, from the digitization of traditional infrastructure (smart grid, water, and mobility), the revolution of industrial sectors (smart autonomous cyber-physical systems, autonomous vehicles, and industry 4.0), to changes in how our society operates (smart government and cities). At the other end of the scale, we see more human-centric thinking in our systems[1] where users have growing expectations for highly personalized digital services for the "market-of-one." The digital transformation is creating a data ecosystem with data on every aspect of our world spread across a range of intelligent systems. Data ecosystems present new challenges to the design of intelligent systems and system of systems requiring a rethink in how we deal with the needs of large-scale data-rich smart environments. How can intelligent systems leverage their data ecosystem to be "smarter?" How can we support data sharing data between smart systems in an ecosystem? How can systems adapt to take advantage of the data within the ecosystem? What are practical approaches to the governance of data within an ecosystem? How can we make trusted decisions using data and humans within the ecosystem? Solving these problems is critical if we are to progress towards next-generation, data-intensive intelligent systems.

FROM DETERMINISTIC TO PROBABILISTIC DECISIONS IN SMART ENVIRONMENTS

Within a smart environment a range of reliability is required. Consider the example of the autonomous connected car/vehicle. We have the strict requirements of safety-critical autonomous driving system, and a failure may lead to loss of life or serious personal injury Compare that to the "good enough" infotainment systems, where a failure is acceptable and merely an inconvenience to the user. When it comes to making decisions in smart systems, there are two general approaches: deterministic (model-driven) and probabilistic (data-driven). A critical difference between the approaches can be explored by considering the costs and level of reliability and adaptability each provides. There is a tension between reliability, predictability, and cost:[1] usually the more dependable and reliable the system needs to be, the more cost is associated with its development. Typically, we can see deterministic systems as reliable but with high costs to develop and adapt (i.e., autopilot), and probabilistic as low-cost to build and adapt, but less reliable (i.e., infotainment).

Where high-levels of reliability are needed, deterministic approaches are an obvious choice for the design of smart systems. This is because the environment is optimized based on a formal deterministic model, and a set of rules and/or equations details the decision logic for the system that is used to control the activity in the environment in an efficient and predictable manner. Adapting the system to meet changes in the environment is a costly process, as the model and its rules need to be updated by expert system engineers.

In the probabilistic approach, the core of the decision process is a statistical model that has been learned from an analysis of training data to learn the structure of a decision model automatically from the observed data (i.e., driver behavior). Thus, a fundamental requirement of data-driven approaches is the need for data to train the algorithms. A lack of data, and training data, within a smart environment limits the use of data-driven approaches.

As the IoT is enabling the deployment of lower-cost sensors, we are seeing broader adoption of intelligent systems and gaining more visibility (and data) into smart environments. Not only are smart environments generating more data, but they are also producing different types of data with an increase in the number of multimedia devices deployed such as vehicle and traffic cameras. The emergence of the Internet of Multimedia Things (IoMT) is resulting in large quantities of high-volume and high-velocity multimedia event streams that need to be processed. The result is a data-rich ecosystem of structured and unstructured data (i.e., images, video, audio, and even text) detailing the smart environment that can be exploited by data-driven techniques.

The increased availability of data has opened the door to the use of the data-driven probabilistic models, and their use within smart environments is becoming increasingly commonplace for "good enough" scenarios. It is estimated that a single connected car will upload about twenty-five gigabytes of data per hour (http://www.cisco.com/web/about/ac79/docs/mfg/Connected-Vehicles_Exec_Summary.pdf), while a vehicle fitted with an autonomous vehicle imaging and scanning system generates and processes about 4 TB of data for every hour of autonomous driving (https://www.datamakespossible.com/evolution-autonomous-vehicle-ecosystem/).

As a result, the conventional rule-based approach is now being augmented with data-driven approaches that support optimizations driven by techniques including machine learning, cognitive, and AI techniques that are opening up new possibilities in the design of smart systems. For example, pedestrian detection is difficult to implement in a rule-based approach. However, deep learning models for object detection and semantic segmentation using a dash-mounted camera are very effective at detecting pedestrians. Systems can now adapt to changes in the environment by leveraging the data generated in the environment within their learning process to improve performance. If systems share data on their operational experiences, then the pooled data can be used to improve the overall learning processes of all the systems, giving us a form of collective artificial intelligence through the "wisdom of the systems." Because the process is data-driven, it can be run and re-run at low cost. This critical role of data in enabling adaptability and collective machine intelligence makes it a precious resource.

SYSTEM OF SYSTEMS

The need to bring together multiple systems within a smart environment to work together is becoming a standard requirement. Initiatives such as Smart Cities are showing how different systems within the city (i.e., energy and transport) can collaborate to maximize the potential to optimize overall city operations. Autonomous connected vehicles can support smart city mobility by providing a vital feedback loop for cities on the state of traffic volumes, flows, roadway design and maintenance, and the mobility requirements (trip information) of its occupants. This requires a System of Systems (SoS) approach to connect systems that cross organizational boundaries (i.e. city, automotive, personal data), come from different domains (i.e., entertainment, manufacturing, logistics, etc.), and operate at different levels (i.e., city, district, neighborhood, fleet, vehicle, or individual passenger). The joint ISO/IEC/IEEE definition of a SoS brings together a set of systems for a task that none of the systems can accomplish on its own. Each constituent system keeps its management, goals, and resources while coordinating within the SoS and adapting to meet SoS goals."[2] Maier[3] identified a set of characteristics to describe a SoS:

- Operational independence: constituent systems can operate independently from the SoS and other systems.
- Managerial independence: constituent systems are managed by different entities.
- Geographic distribution is the degree to which a system is widely spread or localized.
- Evolutionary development: the evolution of a SoS and its behavior, which requires changes to system interfaces to be maintained and kept consistent.
- Emergent behavior: new emergent behavior can be observed when the SoS changes.

There are many challenges in bringing together the constituent systems into a SoS at the data, service, process, and organizational levels that require advanced systems engineering. At the data-level, data-driven approaches can benefit from leveraging data from multiple systems within the smart environment. This requires support for the sharing of data at new scales between multiple complex interconnected system of systems within a smart environment.

DATA ECOSYSTEMS

Within a data ecosystem, participants (individuals or organizations) can create new value that no single participant could achieve by itself.[4] A data ecosystem can form in different ways—around an organization, an activity (mobility), a community of interest (music), a geographical location (city), or within or across industrial sectors (automotive, manufacturing, pharmaceutical). In the context of a smart environment, the data ecosystem metaphor is useful to understand the challenges in maximizing the value of data within the environment. The cross-fertilization and sharing of vital resources and datasets from different participants is a key benefit of data ecosystems, leading to new business opportunities and easier access to knowledge and data.

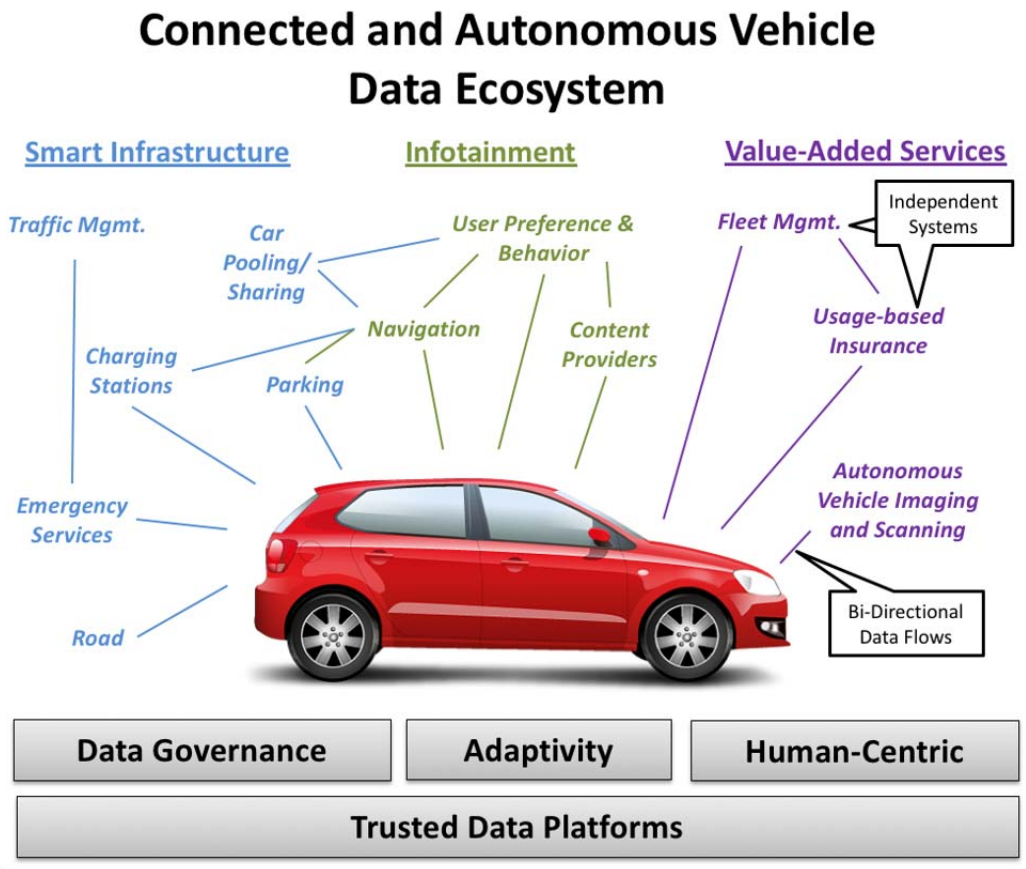

Figure 1. Connected and Autonomous Vehicle Data Ecosystem

Figure 1 details the data ecosystem for connected and autonomous vehicles where a community of interacting data-intensive systems share and combine their data to provide a holistic functional view of the car, passenger, city mobility, and service & infrastructure providers. Systems within the ecosystem can also come together to form a SoS. The ecosystem supports the flow of data between systems, enabling the creation of data value chains to understand, optimize, and reinvent processes that deliver insight to optimize the overall ecosystem. Data may be shared about the current operating conditions of the vehicle, traffic flows, or context of the passengers; a family on holiday, or a business executive moving between meetings. The pooled data can be used to support personalized digital services (i.e. delivering the latest episode of the family's favorite sitcom) and real-time decision-making (i.e. delivering relevant information for the business executive's next meeting). Data on past operating conditions can be shared to improve the learning processes of all systems in the ecosystem.

The nature of the ecosystem, the systems themselves, and the system dynamics will affect the design and operation of the ecosystem. Enabling data ecosystems for smart environments will require a rethink in the design of intelligent systems to consider ecosystem concerns including governance, economics, and technical challenges. Data infrastructure is needed to support data sharing within the ecosystem—from data provided by a single dominant actor on their proprietary infrastructure, to a community pooling their data in a managed open source data platform.

To understand the dynamics of a smart environment data ecosystem we can look to the literature on SoS[3] and business ecosystem[5] to help us understand the different types of data ecosystem that can exist. In Figure 2 we bring together these two areas in the design of a data ecosystem for a smart environment: Koening[5] identified two key criteria regarding the design of a business ecosystem that will also influence a data ecosystem, namely, resource control, and interdependence:

- Control of key data resources: Who controls the essential data resources in the ecosystem? Does a single "keystone"[6] actor control the key data resources that all others depend on, or is control of the key data resources spread across multiple actors in the ecosystem?
- Participant interdependence: Interdependence is based on the degree to which different participants in the ecosystem must interact and exchange data for performing their activities. Reciprocal interdependence requires high levels of coordination between the participants, while pooled interdependence enables loose coupling between participants.

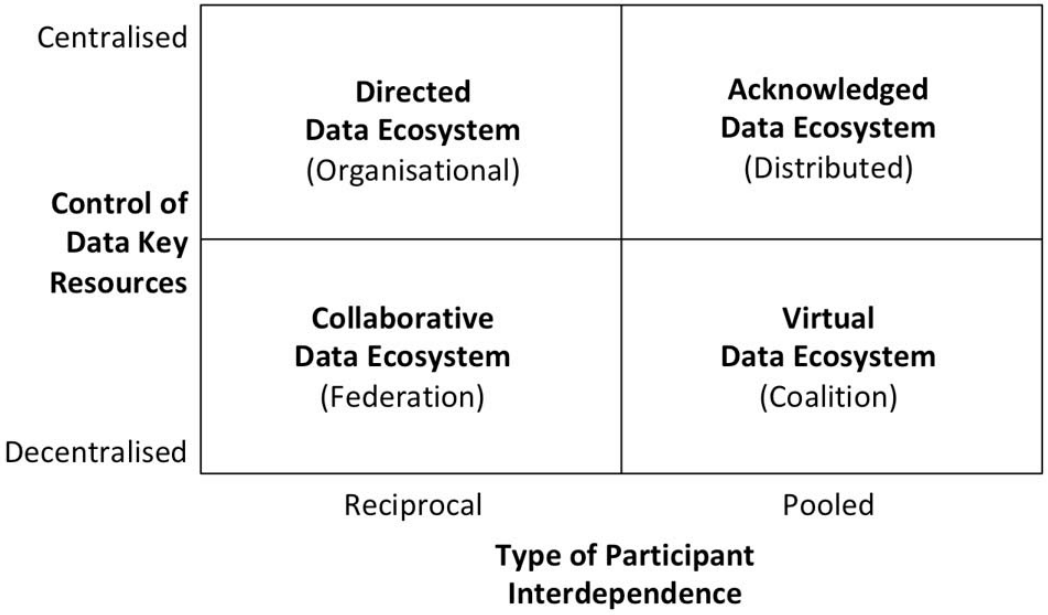

Figure 2. Topology of Data Ecosystems (adapted from Koening[5] and Maier[3])

Drawing inspiration from the SoS classification by Maier[3] (which defines Virtual, Collaborative, Acknowledged, and Directed categories) and the ecosystem topology by Koening, we can consider the different types of data ecosystems that may exist within a smart environment (Figure 2).

- Directed data ecosystems are centrally controlled to fulfill a specific purpose. Typically found within an organization setting or following a keystone model, participants within a directed ecosystem maintain an ability to operate independently, but their standard operational mode is subordinated to the centrally managed purpose of the ecosystem.
- Acknowledged data ecosystems have defined objectives and pooled dedicated resources. The constituent systems retain their independent ownership and objectives. Changes in the ecosystem are based on collaboration between the distributed participants.
- Collaborative data ecosystems have participants interact voluntarily to fulfill an agreed-upon central purpose. The primary players collectively decide the means of enforcing and maintaining standards between the federations of participants.
- Virtual data ecosystems have no central management authority and no centrally agreed upon purpose. Bottom-up coalitions of participants emerge from a virtual data ecosystem to pool decentralized resources to achieve specific goals.

FUTURE DIRECTIONS

Enabling a smart environment data ecosystem will require many challenges to be overcome regarding infrastructure, governance, systems engineering, and human-centricity.

Trusted Data Platforms

To support the ecosystem and the interconnection of systems, there is a need to enable the sharing of data between systems. Platform approaches have proved successful in many areas of technology, and the idea of large-scale "data" platforms have been touted as a possible next step. A data platform focuses on the secure and trusted data sharing among a group of participants (i.e., industrial consortiums sharing private or commercially sensitive data) within a clear legal framework. An ecosystem data platform would have to be infrastructure agnostic and have to support continuous, coordinated data flows, seamlessly moving data between systems. Data exchange could be based on models for monetization or reciprocity. Data platforms can create possibilities for smaller organizations and even individual developers to get access to large volumes of data, enabling them to explore their potential. Data platforms open up many research areas including data discovery, curation, linking, synchronization, standardization, and decentralization. However, the challenges go beyond the technical to issues of data ownership, privacy, business models, and licensing and authorized reuse by third parties.

Ecosystem Data Governance

For mass collaboration to take place within data ecosystems, we need to overcome the challenges of dealing with large-scale agreements between potentially decoupled interacting parties. Research is needed on decentralized data governance models for data ecosystems that support collaboration and fully consider ethical, legal, and privacy concerns. Data governance within an ecosystem must recognize data ownership, sovereignty, and regulation while supporting economic models for the sustainability of the data ecosystem. A range of decentralized governance approaches may guide a data ecosystem from authoritarian to democratic alternatives, including majority voting, reputation models (i.e., eBay), proxy-voting, and dynamic governance (i.e., sociocracy: circles and double linking).[7] Finally, economic concerns may be considered as an incentivization factor within governance models with "data-vote exchange" models where participants pay for votes with data.

Incrementally Evolving Systems Engineering: Cognitive Adaptability

The design of adaptive systems will need to consider the implication of operating within an ecosystem. The boundaries of systems will be fluid and will change and evolve at runtime to adapt to the context of the current situation. However, we must also consider the cost of system participation, and support "pay-as-you-go" approaches at both the system and data-levels. For data management, dataspaces[8] represent one avenue where a pay-as-you-go approach has been applied to integrate data on an "as-needed" basis with the labor-intensive aspects of data integration postponed until they are required.[9] How can the pay-as-you-go approach be extended to the design of incremental and evolving systems?

Work on evolving systems engineering[10] will need to consider the inclusion of data-driven probabilistic techniques that can provide “cognitive adaptability” that will help systems adapt to changes in the environment that were unknown at design-time. Adaptive systems require new iterative development processes that require training and deploying machine learning models over massive volumes of training data with close collaboration between data scientists, software developers, data engineers, and governance professionals. System design will need to consider the varying levels of accuracy offered by data-driven approaches, providing best-effort or approximate results using the data accessible at the time.[8] How can we mix deterministic and statistical approaches? How can we test and verify these systems? What are the challenges in making decisions using multiple sources from the ecosystem?

Towards Human-Centric Systems

Currently, intelligent systems make critical decisions in highly-engineered systems (i.e., autopilots) where users receive specialized training to interact with them (i.e., pilots). As we move forward, intelligent systems will be making both critical and lifestyle decisions—from the course of treatment for a critical illness and safely driving a car, to choosing what takeout to order and the temperature of our shower. Data-driven decision approaches (including cognitive and AI-based techniques) will need to provide explanations and evidence to support their decisions and guarantees for the decisions they recommend. The role of users in data ecosystems will not be a passive one. Users are a critical part of socio-technical systems, and we need to consider more ways of including the "human in the loop" within future systems. Active participation of users can improve their engagement and sense of ownership of the system. Indeed, active involvement of the user could be a condition for them granting access to their private data. Research is needed to build trust in algorithms and data—in the trusted co-evolution between humans and AI-based systems, and in the legal, ethical, and privacy issues associated with making data-driven critical decisions.

ACKNOWLEDGEMENTS

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie).

REFERENCES

1. A. Sheth, "Computing for Human Experience: Semantics-Empowered Sensors, Services, and Social Computing on the Ubiquitous Web," *IEEE Internet Computing*, vol. 14, no. 1, 2010, pp. 88–91.
2. *ISO/IEC/IEEE 15288: 2015 Systems and Software Engineering - System Life Cycle Processes*, standard ISO/IEC/IEEE 15288, ISO/IEC/IEEE, 2015.
3. M. W. Maier, "Architecting Principles for Systems-of-Systems," *Systems Engineering*, Wiley, 1998.
4. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, J. M. Cavanillas, E. Curry, and W. Wahlster, Springer International Publishing, 2016.
5. G. Koenig, "Business Ecosystems Revisited," *Management*, vol. 15, 2012, pp. 208–224.
6. H. Kim, J.-N. Lee, and J. Han, "The Role of IT in Business Ecosystems," *Communications of the ACM*, vol. 53, 2010, p. 151.
7. J. A. Buck and S. Villines, *We the People : Consenting to a Deeper Democracy : a Guide to Sociocratic Principles and Methods*, Sociocracy.info, 2007.
8. M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," *ACM SIGMOD Record*, vol. 34, no. 4, 2005, pp. 27–33.
9. E. Curry et al., "Internet of Things Enhanced User Experience for Smart Water and Energy Management," *IEEE Internet Computing*, vol. 22, no. 1, 2018.
10. M. Hinchey and L. Coyle, "Evolving Critical Systems: A Research Agenda for Computer-Based Systems," *217th IEEE International Conference and Workshops on Engineering of Computer Based Systems*, 2010, pp. 430–435.

ABOUT THE AUTHORS

Edward Curry is a funded investigator at Lero: The Irish Software Research Centre and a lecturer in informatics at the National University of Ireland Galway. http://edwardcurry.org

Amit Sheth is the LexisNexis Ohio Eminent Scholar and the executive director of Kno.e.sis - Ohio Center of Excellence in Knowledge-enabled Computing and BioHealth Innovations. He is a fellow of the IEEE and the AAAI. http://knoesis.org/amit