Pervasive Data Science

DOI:
10.1109/MPRV.2017.2940956

Document Version
Accepted author manuscript

Pervasive Data Science: New Challenges at the Intersection of Data Science and Pervasive Computing

Nigel Davies and Sarah Clinch

Abstract—Technology is increasingly enabling us to instrument our physical environment with complex sensors and actuators, creating a connected world that generates huge volumes of complex data. In this article we describe a number of opportunities and challenges that this new data-rich world brings to pervasive computing and highlight the emergence of a new field of pervasive data science.

1 Introduction

Recent years have seen the rise of data as a central tenant in computing applications, products, research, and innovation. Commentators have identified a big data "paradigm shift" [17], and use of the term "data science" to describe the interdisciplinary field of collecting, drawing inference from and acting on data has grown exponentially (e.g. Google Trends reveals a ten-fold interest in search volume for the term since 2010). Beyond the hype, it is clear to see how our world is genuinely becoming one that is increasingly data-centric, in which both physical and electronic services depend on the collection, analysis, and response to, large volumes of heterogenous data.

Examples of the success of data science abound – the application of new machine learning techniques to problems such as speech recognition have made common place levels of performance that would have seemed impossible a few years ago. Apple's Siri is reported to receive over 2 billion requests per week and provides clear evidence of the transformative nature of such innovation. Such successes have encouraged substantial levels of new research into algorithms and systems for data processing, storage and visualisation.

In contrast to this focus on data, pervasive computing has historically often focused on user experience, principally motivated by Weiser's original papers and his compelling vision of "calm computing" [18]. Of course data has often had a role to play in providing this user experience, and research areas such as activity recognition and location technologies are highly quantitive in nature. However, in recent years the field of pervasive computing has invested significant effort into understanding the societal implications and applications of Weiser's vision. This focus on the human elements of pervasive computing is in contrast to technology and application trends that increasingly enable us to instrument our physical environment with complex sensors and actuators, creating a connected world that generates huge volumes of interconnected data. The importance of these trends can be seen in the growing momentum of exemplars such as the Internet of Things (IoT), smart environments and smart cities. These applications demand a new focus on how we capture, process and utilise data in pervasive environments.

In this article we identify a new field of research that we term *pervasive data science*. Our objective is to highlight the importance of work in this area and identify examples of key research challenges for the community, thus acting as a catalyst for new research. We define pervasive data science as research that exists at the intersection of pervasive computing and data science and is characterised by *a focus on the collection, analysis (inference) and use of data (actuation) in pursuit of the vision of ubiquitous computing*. Examples of topics that we would consider part of the pervasive data science agenda include the design of new IoT sensor networks, new architectures for data processing at the edge of cloud, algorithms for processing pervasive sensor data, new techniques for data visualisation in pervasive environments and cross-cutting concerns such as privacy and trust. These topics highlight the inherently multidisciplinary nature of the field of pervasive data science – we are not simply talking about a shift towards more quantitive ubicomp research but a more fundamental refocusing on data as a key element in delivering Weiser's vision.

2 Example Application Areas

To what extent does pervasive data science really have an important role to play in our everyday lives? Understanding the breadth of data-driven pervasive applications is a worthwhile foundation for defining the field. In the following examples we illustrate some of the potential opportunities for data-driven pervasive applications. While the application areas may not be completely new, to date pervasive computing has failed to make them an everyday reality – we believe that the new capabilities provided by pervasive data science will be transformative in enabling such applications.

2.1 Augmented Cognition

Using pervasive technology to improve human's cognitive capabilities. Our everyday lives feature a huge range of cognitive activities, and technologies and artefacts have long-played a role in supporting these – we write lists to evaluate the pros and cons of an important decision; take photographs of experiences we want to remember, and record voice memos

to help with productivity. As data collection, processing, and presentation becomes all-encompassing, the potential for more effectively supporting these processes increases dramatically. Indeed, since many have proclaimed that industrialisation, technology, etc. have greatly increased the cognitive demands on human beings [16], it seems only fitting that technology now helps to address these very concerns by helping to support cognition and reduce load. The exact nature of these technologies in terms of user interfacing is currently unclear, nevertheless there is undoubtedly an important role for data-intensive pervasive systems in a huge array of cognitive processes including decision making; evaluation of risk; regulating mood and emotion; creative thinking; attention and information processing; and retrospective and prospective memory.

Designing these systems will require a clear understanding of cognition. Psychologists have recently begun to distinguish between two metaphorical systems of thought – System 1 ("intuition") and System 2 ("reasoning") [9]. While operations of System 1 are fast, automatic and effortless, they are also often emotionally charged, governed by biases and habit, associative, and difficult to control. By contrast, the operations of System 2 are slower, effortful, resource-intensive, and deliberately controlled, but are are also more adaptive and neutral with regard to emotion and bias. Pioneered as a metaphor for psychologists and behavioural economists, this distinction may also be valuable in understanding target processes to support with pervasive data applications. Critically, System 1 and 2 thinking is often used to explain flaws in our cognition – e.g. in the evaluation of risk, decision making – as being primarily a result of relying on use of System 1 thinking because we have either insufficient time or inclination to engage in deeper processing. If technology can reduce the barriers to System 2 thinking, or indeed provide System 2 like thought in System 1 timeframes, then many of the impairments to reasoning caused by System 1's inherent biases may well be overcome. For example, one could easily envisage data processing techniques being put to use to reduce the effort of System 2 thinking, whilst the additional data provided by pervasive sensing could also provide a richer input to that processing.

In addition to helping address flaws in cognition generally, pervasive data science could help to address growing concerns with regard to mental health problems and disorders of cognition. In 2015, the US National Institute of Mental Health[1] estimated the prevalence of diagnosable mental disorder at 1 in 4 adults, with nearly 1 in 25 having serious functional impairment due to a mental illness. In terms of cost, it is estimated that this amounts to in excess of $300 billion per year. Similarly, the rising prevalence of dementia incurs costs of around $200 billion annually, more than both heart disease and cancer, largely due to the need to provide institutional and home-based long-term care for individuals that are no longer able to engage in their day-to-day activities without support. Providing unobtrusive support that helps individuals maintain their own identities, mood, emotion, attention, memories and thought processes could be critical in reducing the costs associated with cognition and mental health related illness and decline.

1. https://www.nimh.nih.gov/

2.2 Autonomous Vehicles

Delivering the vision of mobility as a service.

Ever since Weiser's original description of the "foreview mirror" pervasive computing has looked to enable smart transportation; autonomous vehicles represent a natural extension of this work. For many years, computation and electronics have been an increasing feature of cars and other vehicles. However, the wealth of available sensor data (both in the environment and on-board the vehicle) together with advances in machine learning to interpret the data and predict future environmental changes means that we have now moved from *automation* to genuinely *autonomous* vehicles. The arguments for such vehicles are well rehearsed: improved safety is perhaps the most obvious, eco-driving is another, as is the reduced burden for users.

Although one could simply envisage autonomous vehicles as enhanced cars that improve each individual journey by allowing those being transported to travel more comfortably, safely or economically, the reality is that they will offer new opportunities for transport and logistics as a whole – forming part of a new vision of "mobility as a service". This vision extends beyond autonomous vehicles and imagines a world in which mobile applications allow travellers to dynamically select the most appropriate form of transportation to achieve their mobility objectives with potentially profound implications for transport infrastructure, urban planning and economics.

2.3 Smart Spaces

Weaving technology into the fabric of our everyday lives to improve our physical environment.

Exemplars such as the Internet of Things (IoT), smart environments and smart cities are arguably the most obvious current applications of pervasive data science. Technology increasingly enables us to instrument our environment with sensors and actuators, creating a connected world that generates huge volumes of complex data.Of course, the addition of 'smart' into our environments is not new for pervasive computing – Weiser's vision for Ubicomp included domestic appliances that interpreted future needs based on recent activity, neighbourhoods that tracked mobility patterns, and an office that supported awareness and communication between remote colleagues. However, progress in data science is beginning to provide tools with which to realise elements of this vision and to offer previously unenvisioned services based on the collection, analysis and application of data in our physical environments.

Applying pervasive data science to our physical environments offers a wealth of opportunities for improving quality of life through access to smarter and more appropriate services. Whilst some of these advantages span across almost all of our physical spaces, there are also a number of specific goals that can be addressed in particular target environments. For example, we can use pervasive data to make our workspaces more pleasant and efficient, answering questions such as: what physical conditions lead to a satisfied and productive workforce? how do we best foster inter-team collaboration? which portion of a workflow

offers the most potential for optimisation?. Schools and places of education can build an understanding of how non-classroom environments (corridors, social spaces) can be used to complement learning. Similarly, use of pervasive data science in outdoor environments can help tackle challenges like climate change, traffic congestion and urban air pollution.

While many of the above examples approach smart spaces as something that benefits populations and organisations as a whole, pervasive data science also lends itself perfectly to tailoring of physical environments in order to provide a personalised experience for the individual. Shared spaces could be uniquely configured to respond the user within them, changing not only the aesthetics but also the physical configuration of the space itself.

3 Challenges

3.1 Overview

In considering the challenges associated with pervasive data science it is worth reflecting on those traditionally associated with its two parent domains of data science and pervasive computing. While data science is a relatively new field it draws on many years of prior work in areas such as statistics, algorithms and databases. Data science challenges are often based on the characteristics of so-called big-data, as expressed in the "3Vs" of i.e. Volume, Variety and Velocity [11] and sometimes supplemented with an additional V in the form of Veracity (we revisit these challenges in Section 3.3), though it is worth noting that general papers on the challenges of data science are rare.

By contrast, there are a large number of papers that have articulated challenges relating to mobile and ubiquitous computing, including [14] and Weiser's original paper [18]. Indeed, the inaugural issue of IEEE Pervasive Computing focused on articulating the challenges that still existed in the field ten years after its inception. Many of these challenges remain and cover a broad space including the development of appropriate systems architectures, new forms of user interaction and cross-cutting concerns such as ease of deployment and system maintainability.

In this section we set out some of the new challenges that arise at the intersection of these two fields. We structure our discussion in terms of the key stages in any data processing application, i.e. data collection, inference and subsequent action.

3.2 Data Collection

The first stage in any data pipeline is data collection (and cleaning). Pervasive data can originate from a wide range of sources including sensors embedded in the environment, sensors attached to users, and explicit user input as in the case of initiatives such as citizen science. While the design of any large-scale data collection system is non-trivial, the emergence of systems to support pervasive data science give rise to a number of exciting new research challenges.

How do we manage complex models of data and sensor ownership? In conventional sensor systems the question of data ownership is relatively clear – the data owner is normally the same as the owner of the system being instrumented. Similarly, most data scientists assume data ownership has been resolved prior to them obtaining data for analysis. However, in a world of pervasive sensors the question of data ownership becomes significantly more complex. For example, in many smart environment applications the same space may be instrumented by many different stakeholders, mobile users may bring their own sensors, or wish to use those of their peers, and the use of spaces and sensors may be highly transient. As a result the ownership of any given data stream (or combination thereof) may be unclear. Can we develop techniques for automatically resolving data ownership in such pervasive computing scenarios? How do we model shared ownership of data? How do we accommodate ownership expectations when considering personal data that users typically perceive as belonging to them even though they may not own the sensing infrastructure (e.g. information on energy and activities in a domestic environment collected by a smart heating management system)? How do we ensure that ownership of data reflects transient use of pervasive sensors and spaces?

Can data provenance be ensured in pervasive data systems? Determining the provenance of data in existing systems is a well documented research challenge with solutions typically involving techniques such as audit trails based on digital signatures. However, in pervasive systems many new aspects of data provenance become important [10]. For example, how do we capture the identity and motivations of humans involved in sensor placement given that even a small bias in the placement of sensors may have a significant impact on the data captured? Once provenance data has been captured, how should this be presented to end-users to enable them to understand the likely impact on the data? Given that pervasive systems are likely to involve complex data pipelines it makes sense to record details of these pipelines and yet some aspects (e.g. data redaction policies) may be sensitive. How do we balance the need for end-to-end provenance with the need to mask the identity and motivations of some users?

How do we resolve the tension between pervasive data science's insatiable demand for data and concerns regarding user privacy? Privacy has long been recognised as a challenge in pervasive environments [12]. Despite extensive research, the challenge of protecting user privacy remains and, indeed, is becoming significantly more difficult to address as the number of sensors in the environment increases. Are we now reaching the point where the only way to effectively protect user privacy is to limit data collection at source? If so, new architectural solutions will be required to enable data to be quenched at source. In [6] the authors propose a model in which users are able to control the release of data from their homes. While the model is simple "users should be able to control the release of their own data" the implementation is complex, necessitating the introduction of new architectural components such as privacy mediators that are able to denature data prior to disclosure [6].

How will data subjects provide informed consent? Related to, but distinct from, the issue of user privacy is the challenge of supporting informed consent. How should we inform users of data collection in a pervasive data environment? It is clearly impractical to explicitly prompt users every time they enter an environment in which data is being captured

about them. However, it is equally important to ensure that users understand and consent to the collection of information and, crucially have a genuine option to opt-out of data collection. Once consent has been provided many systems provide tools for data collectors to track subject consent. However, there is a dearth of tools that enable users themselves to track when and where they provided consent. This opens up the opportunity for unscrupulous data collectors to simply claim that consent has been provided – how many users could really assert with confidence whether or not such a claim was true? While this discussion suggests that pervasive data science is likely to make managing consent extremely challenging, pervasive data science also offers the potential for entirely new ways of managing consent – based on systems that automatically learn user's preferences and behaviours and infer whether or not to automatically provide consent.

3.3 Inference

Inference lies at the heart of what many consider to be data science and represents the process of analysing data to gain understanding and insight.

How does pervasive computing impact on traditional challenges of data science? In future pervasive environments the *volume* of data is likely to dwarf that produced by most existing data systems – for example, while classic data systems may examine feeds such as web browsing histories or social media posts, pervasive data applications operate in a world in which every aspect of a user's experience is captured by sensors. As a point of reference, current life logging cameras capture over 2000 images per day – far exceeding the number any user might manually process or post in a typical day. How can we store and process such volumes of data? Widespread user and environmental sensing is also likely to lead to a *variety* of data previously unseen – creating heterogenous datasets in terms of format, frequency and quality (amongst others). This raises challenges in terms of data consistency but opens up exciting new possibilities – e.g. how can we effectively combine data from such a wide range of sensors? While any individual sensor is unlikely to produce very high-velocity data, the number of sensors in any given environment is likely to lead to data aggregators experiencing streaming data at unprecedented *velocity*. This raises important questions regarding future data processing architectures.

How should we architect pervasive data science environments? Existing data systems tend to assume large-scale data centres that are available to carry out the processing necessary to draw inference. Where sensors are the source of this data they are typically assumed to be "dumb" sensors with all of the processing being conducted in the cloud. However, research has clearly demonstrated the shortcomings of a purely cloud-based model [15] and new architectures have been proposed that provide data processing at the edge of the cloud. What are the correct processing architectures for future pervasive data environments? Future pervasive environments may offer a wide range of options for hosting data processing – from highly sophisticated sensors that can carry out significant levels of analysis on-board, through edge-of-cloud solutions to micro and and full data centres. To what extent does the exact configuration of processing elements used depend on the intended applications and how will it be influenced by security and privacy concerns?

3.4 Actuation

Traditional data science has often focused on information presentation as its key output. Likewise, other fields that work with large emerging datasets (e.g., computational social science [13]) reach their end with the analysis of data representing human behaviour patterns. Each of these fields achieve actuation only through humans implementing changes based on the outcome of data. By contrast, actuation has been a common feature of ubiquitous computing since Weiser's early vision, and exemplars such as the IoT and smart cities continue to demonstrate its importance. While, pervasive data science requires inference to deliver value, its data lifecycle does not end with analysis; instead pervasive data applications offer opportunities and challenges in both new forms of information presentation and physical actuation.

How can we best use pervasive technology to help visualise rich data sets? Data scientists have always sought new ways of visualising data. Pervasiveness provides a wealth of new presentation opportunities, enabling the process of engaging with data to be switched from one of active interpretation to one of passive immersion. Weiser observed that natural environments can convey a wealth of information that can be readily absorbed and yet still deliver a positive user experience, and a number of trends in pervasive displays are helping to realise a similar paradigm for digital interactions [4]. Given the range of available technologies, how do we select the most appropriate medium for engaging users? How do we take into account factors including the contextual relevance of the device, the scale and resolution of data that can be represented, the shared or private nature of the content, and individual aspects such as attention and task engagement? While data science has focused primarily on visualisation of data for expert users, pervasive data potentially creates a requirement for information presentation to become accessible to a wide range of individuals. This "data-for-the-masses" poses considerable new challenges – how do we develop a set of patterns for comprehendible representations (not necessarily visual) that are accessible to populations of different ages, education levels and cultural backgrounds?

What new forms of data-driven actuation will emerge? Visualisation is just output form for pervasive data and numerous other forms of data-driven actuation exist – for example, data-driven control and adjustment of our environments; such adjustment may take the form of a slow transition that optimises performance of a space, or a more dynamic process that personalises environments to the changing presence of individuals and purpose. What new possibilities for smart-spaces will emerge with widespread data availability? How will users be made aware of such data use, and how will they exercise control over actuation? Of course data-driven smart spaces are just one example and numerous other forms of data-driven actuation are likely to emerge in areas such as the IoT.

3.5 Pervasive Data Science in Context

The questions raised in this section highlight the research opportunities that exist at the intersection of data science and pervasive computing. Pervasive computing brings new challenges to conventional data science tools by virtue of both the scale at which it operates and its focus on the relationship between technology and users. However, the successful application of data-driven approaches to pervasive computing challenges open up the possibility of genuinely smart environments and applications as described in section 2.

Of course there have been numerous attempts to transfer the techniques and insights of data science to a broad set of disciplines. For example, almost ten years ago, Lazer et al. [13] described the emergence of data-driven computational social science – an application of data at scale to describe individual and group interactions. Likewise, social and community intelligence (as articulated by Zhang, Guo and Yu in 2011 [19]) leverages these same tools to reveal human behaviour patterns and dynamics. Furthermore, ubicomp itself has often had a data dimension (e.g. the computational location applications of John Krumm and others [7]), incorporating qualitative and quantitative data from a range of sources to capture both 'the masses' and the individual. While domains such as computational social science and quantitive ubicomp have focused predominantly on data collection and analysis, one unique feature of pervasive data science is its end-to-end use of data – collection, analysis *and* actuation. Indeed, we believe that the emerging trend for pervasive data science is the natural progression of such disciplines, but broadened to reflect the original ambition of Weiser's ubiquitous computing.

Another key distinction of pervasive data science is its interdisciplinary nature – whilst a field such as computational social science integrates computing methods with those of the social sciences, pervasive data science must bring together social sciences and humanities with a broad set of computer science expertise – including theory, systems architecture, human-computer interaction. The challenges that we have highlighted all span multiple fields of computer science (see Figure 1) and this is deliberate – we believe that this intersection offers new opportunities for interdisciplinary research and development, and is a distinct feature of this emerging field.

	Theory	Systems	People
Collection	Provenance and Ownership		
	Algorithms for signing streaming data that are optimised for use on low-power sensors.	End-to-end secure architectures and protocols for high velocity data streams.	Techniques for presenting provenance to users. Models of data ownership.
	Privacy and Consent		
	New algorithms for data denaturing.	System architectures that encompass privacy by design.	New UI techniques for obtaining informed consent.
Inference	Challenges to Traditional Data Science		
	Techniques for compensating for highly variable and low-quality data.	Storage systems for high volume, high velocity data.	Tools for helping users understand inference's drawn from their data.
	New Architectures		
	Algorithms optimised for distributed operation at the edge of the cloud.	Novel architectures to support computational off-loading.	What are acceptable business models for data processing and service provision.
Actuation	Pervasive Technology for Information Visualisation		
	New algorithms for info-viz on distributed displays.	Architectures for display coordination.	Guidelines for preventing information overload in pervasive environments.
	New forms of Data-drive Actuation		
	Formal models of how environments respond to different data inputs.	Protocols for secure communication with IoT actuators.	Interfaces for user-control of data use.

Fig. 1. Examples of emerging challenges in pervasive data science

4 Early Case Studies

As a prime example of these interdisciplinary challenges in practice, we return to our earlier scenarios. Specifically, we consider two specific case studies of work in the area of augmented cognition, specifically on task assistance and augmented memory. These examples provide an important illustration of the end-to-end nature of pervasive data science, whilst providing a foundation from which challenge areas may be considered and addressed.

4.1 Cognitive Assistance for Task Completion and Decision Support

Reseachers at CMU have set out a vision for *angel on your shoulder* cognitive assistance that takes the familiar models of detailed directions given by current GPS systems and applies them to everyday living [3]. Such systems combine wearable sensing and presentation with local processing in order to guide users through complex tasks, telling them what to do next (with audio-visual cues) and correcting any erroneous actions. Whilst arguably an ambitious concept with very broad parameters that apply in virtually all facets of everyday life, the Carnegie Mellon-based group have focused predominantly on support for well-defined tasks that may require the prompt application of specific knowledge or skills (e.g. administering first aid, engaging in competitive sport). A key challenge is that completion of such tasks is both latency sensitive (a decision must be made in finite time) and resource intensive (requiring processing of the current context, identification of relevant skills or knowledge, and determination of the correct behavioural response).

To address these challenges, the Gabriel platform [3] uses a cloudlet computing infrastructure [15] in which each cognitive process (e.g. face recognition, motion classification) is encapsulated in its own virtual machine (VM). Each VM is then able to independently process incoming sensor data from a control VM. Any resulting output from the cognitive processing is then passed back to a shared user guidance VM which can aggregate the results and perform any high-level integration or further processing needed to deliver assistive input to the user. The Gabriel architecture has been verified with an initial set of demonstrators indicating that cloudlets could be a key part of future cognitive assistance applications.

The assistive output provided in Gabriel and its associated demonstrators are a clear illustration of pervasive data science in practice. Each of the four demonstrator applications relied heavily on worn sensors to capture rich

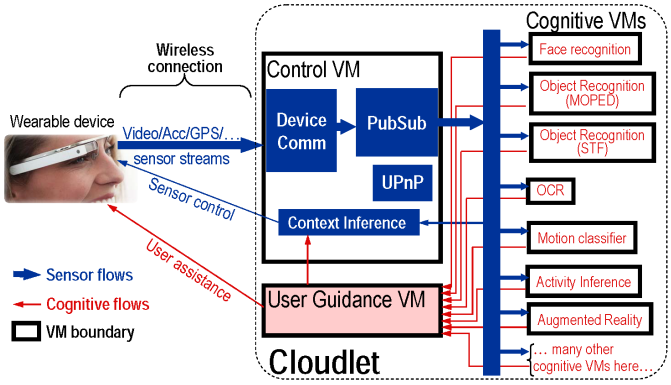

Fig. 2. The Gabriel Architecture for Cognitive Assistance [3]. Used with Permission.

(typically visual) data feeds representing the task to be completed and the associated context. Computer vision and other data processing techniques can then be used to interpret sensor input and predict the outcome of multiple future behaviours, leading to an optimal target behaviour that the assistant should support the user in achieving. Further processing of historical data may then be needed to identify the best cues and presentation medium for this user in the given context. Finally, the system then presents input to the user and continues collecting data to close the loop and measure the ongoing success of its advice.

4.2 Human Memory Augmentation

Our own research has seen us address human cognition in terms of retrospective and prospective memory. Human memory is critical to self-identity, and to the success of most of our everyday activities. Whilst prospective memory (memory for intention) is a relatively constrained problem space, retrospective memory encompasses recollection of personal experiences (episodic memory), knowledge acquired (semantic memory), and motor skills learned (procedural memory). With this in mind, we focus here predominantly on episodic memories (although many of our approaches may well generalise beyond these).

Through a series of prototypes and deployments, our research has led us to an architecture that combines mobile, wearable and environmental devices to capture a rich representation of an individual's experience (and thus derived knowledge) [5], including occurrences in which the human memory itself was known to have failed. These data streams serve as input to a graph-based storage, the *memory vault*, which attempts to reflect human cognition in its model of both how individual pieces of information connect together to form a single digital memory, and also the interconnection between multiple digital memories. Since many human experiences are the product of social interactions, and since memory of those experiences is continuously shaped by ongoing interaction with those (and other) individuals, it is not the case that a single centralised digital memory will meet an individual's needs. Instead, we anticipate that our memory vault will be a distributed memory, with interconnections spanning multiple instances, and with data from others being made readily available to those who shared an experience [2].

Human memories continuously draw on traces formed over many years, and ongoing processing and inference are clearly critical to the success of memory augmentation systems. Unlike the example of task assistance that focused primarily on real-time processing, our memory augmentation architecture makes heavy use of continuous long-term data aggregation and inference; the output of these is stored alongside raw experience data in the memory vault. When an individual then engages in a situation that would be facilitated by access to extended memory, our architecture can pass current contextual feeds to the memory vault and select an appropriate cue for presentation. The system then presents input to the user and continues collecting data to close the loop and measure its ongoing success.

In line with challenges outlined earlier, our research in the area of augmented memory has highlighted significant security and privacy concerns related to the acquisition, processing and presentation of datasets [5]. Bystander privacy is one that is often raised (and is thus the target of considerable research to date [1], [8]), but many more subtle data challenges also emerge. For example, the issue of provenance discussed above becomes important in ensuring that digital memory is an accurate representation of what occurred, particularly since the very need for augmented memory means the human may be unable to determine this for themselves. The phenomenon of recall induced forgetting is a good example of a threat that arises only when our understanding of human cognition is applied to the design of data-intensive pervasive applications – in this case, psychology theory indicates that rehearsal of one memory has a detrimental effect on the subsequent recall of related memories, a clear concern for those designing technology interventions in this space.

5 Concluding Remarks

Recent developments in fields such as machine learning have demonstrated the potential of data science to transform our ability to deliver solutions to traditionally very demanding problems such as speech recognition. These developments, coupled with widespread deployment of sensing and actuating technologies mean that there is the potential for pervasive computing to adopt an increasingly data-driven approach – which we have termed pervasive data science.

In analysing the challenges such an approach raises we observe that many are cross-cutting in two distinct axis. Firstly, the challenges span all stages of the data pipeline – from data collection through to data visualisation and actuation. Secondly, the challenges require multi-disciplinary approaches as they raise issues in terms of systems, algorithms and people. For example, protecting user privacy in future pervasive environments demands new systems work on topics such as privacy mediation, new algorithms for efficient data denaturing, and new techniques to enable users to visualise and control the release of their personal data. However, the pay-off for developing solutions to these problems is significant – applications such as augmented cognition and memory have the potential to transform our society, delivering benefits to millions.

Cross-cutting, multi-disciplinary problems are, of course, common in traditional pervasive computing research and

we therefore believe that the field is particularly well suited to addressing these types of problem. Pervasive computing has always favoured researchers that adopt a holistic view, and the emergence of pervasive data science serves to reenforce the validity of this approach.

ACKNOWLEDGMENTS

This research is partially funded through the Future and Emerging Technologies (FET) programme within the 7th Framework Programme for Research of the European Commission, under FET grant number: 612933 (RECALL) and the EPSRC under grant EP/N028228/1 (PACTMAN:Trust, Privacy and Consent in Future Pervasive Environments).

REFERENCES

[1] P. Aditya, R. Sen, P. Druschel, S. J. Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. Wu. I-pic: A platform for privacy-compliant image capture. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, pages 235–248, New York, NY, USA, 2016. ACM.

[2] A. Bexheti, M. Langheinrich, and S. Clinch. Secure personal memory-sharing with co-located people and places. In *Proceedings of the 6th International Conference on the Internet of Things*, IoT 2016, 2016.

[3] Z. Chen, L. Jiang, W. Hu, K. Ha, B. Amos, P. Pillai, A. Hauptmann, and M. Satyanarayanan. Early implementation experience with wearable cognitive assistance applications. In *Proceedings of the 2015 Workshop on Wearable Systems and Applications*, WearSys '15, pages 33–38, New York, NY, USA, 2015. ACM.

[4] S. Clinch, J. Alexander, and S. Gehring. Pervasive displays for information presentation. *Pervasive*, 15(3), 2016.

[5] N. Davies, A. Friday, S. Clinch, C. Sas, M. Langheinrich, G. Ward, and A. Schmidt. Security and privacy implications of pervasive memory augmentation. *Pervasive Computing, IEEE*, 14:44 – 53, 2015.

[6] N. Davies, N. Taft, M. Satyanarayanan, S. Clinch, and B. Amos. Privacy mediators: Helping IoT cross the chasm. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, HotMobile '16, pages 39–44, New York, NY, USA, 2016. ACM.

[7] D. Delling, A. V. Goldberg, M. Goldszmidt, J. Krumm, K. Talwar, and R. F. Werneck. Navigation made personal: Inferring driving preferences from gps traces. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 31. ACM, 2015.

[8] M. Fan, A. T. Adams, and K. N. Truong. Public restroom detection on mobile phone via active probing. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, ISWC '14, pages 27–34, New York, NY, USA, 2014. ACM.

[9] D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):1449–1475, 2003.

[10] B. Knowles. Emerging trust implications of data-rich systems. *Pervasive*, 15(4):76–84, 2016.

[11] D. Laney. 3d data management: Controlling data volume, velocity and variety. Technical report, Gartner, 2001.

[12] M. Langheinrich. A privacy awareness system for ubiquitous computing environments. In *Proceedings of the 4th International Conference on Ubiquitous Computing*, UbiComp '02, pages 237–245, London, UK, UK, 2002. Springer-Verlag.

[13] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[14] M. Satyanarayanan. Fundamental challenges in mobile computing. In *Proceedings of the Fifteenth ACM Symposium on Principles of Distributed Computing*, 1996.

[15] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies. The case for vm-based cloudlets in mobile computing. *Pervasive*, 8(4), 2008.

[16] C. S. Saunders and A.-F. Rutkowski. Growing pains with information overload. *Computer*, 43, 2010.

[17] R. Schutt and C. O'Neil. *Doing Data Science: Straight Talk from the Frontline*. " O'Reilly Media, Inc.", 2013.

[18] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–104, 1991.

[19] D. Zhang, B. Guo, and Z. Yu. The emergence of social and community intelligence. *Computer*, 44(7):21–28, 2011.

Nigel Davies is a Professor of Computer Science at Lancaster University, UK, and co-director of Lancaster's Data Science Institute. His research is in the area of pervasive computing including systems support for new forms of data capture and interaction and is characterized by an experimental approach involving large-scale deployments of novel systems with end-users. Contact him at n.a.davies@lancaster.ac.uk.

Sarah Clinch Dr. Sarah Clinch is a computer science researcher and lecturer in the University of Manchester, UK. She has a PhD in Computer Science from Lancaster University. Sarahs research interests include applications for human cognition, pervasive display deployments, and privacy and personalisation in ubiquitous computing systems. Sarah is an inaugural member of the ACM Future of Computing Academy (ACM-FCA). Contact her at sarah.clinch@manchester.ac.uk.