Key Challenges for Multimedia Research in the Next Ten Years

March 2019

Shih-Fu Chang, Columbia University
Alex Hauptmann, Carnegie Mellon University
Louis-Philippe Morency, Carnegie Mellon University
Sharon Oviatt, Monash University
Hari Sundaram, University of Illinois at Urbana-Champaign

This short paper is based on summary of the full report [1] of the NSF Workshop on Multimedia Challenges, Opportunities and Research Roadmaps held on March 30-31, 2017 in Washington DC. This material is based upon work supported by the National Science Foundation under Grant No. 1735591. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The workshop participants include the following.

2017 NSF Multimedia Workshop Participants:

Terry Adams
Sameer Antani
Dick Bulterman
Carlos Busso
Shih-Fu Chang
Joyce Chai
Alex Hauptmann
Julia Hirschberg
Ramesh Jain
Ketan Mayer-Patel
Reuven Meth
Raymond Mooney
Louis-Philippe Morency
Klara Nahrstedt
Shri Narayanan
Prem Natarajan
Sharon Oviatt
Balakrishnan Prabhakaran
Arnold Smeulders
Hari Sundaram
Adam Wolfe
Zhengyou Zhang
Michelle Zhou

1. Introduction

The multimedia and multimodal (MM) research community is dedicated to research on the acquisition, communication, analysis, modeling, interface design, and impact of applying multimodal-multimedia data to challenging problems such as semantic information extraction, co-processing and interpretation of multiple heterogeneous modalities, seamless human-machine interaction, and scalable distributed collaboration. This community of researchers also aims to advance important applications such as education, healthcare, advanced communications and social networking. One of the key objectives for MM research is to model and, in some cases, exceed humans' ability to process multi-sensory information during activities like communication and learning. The multidisciplinary MM field is distinct from other research areas in its emphasis on deep, meaningful integration of multimodal data to enhance the reliability and depth of information derived, as well as the quality of multimedia user experience. A major advantage of fusion-based multimodal interaction and systems is greater robustness and coverage of information processing and improved experience of interaction, which typically exceed that possible based on the sum of individual modalities. And, the synergy between multimedia-multimodal data sources is often the prime source of generating new knowledge.

The MM community has demonstrated extremely productive activities in the last two to three decades. Starting with the inaugural ACM Conference in Multimedia (ACMMM) in 1993, the ACM International Conference on Multimodal Interfaces (ICMI) in 2003, and high-quality publications in other professional societies, the MM community has celebrated major technological breakthroughs with a major impact on industry and society. Within the past decade, the dominant computer interface paradigm worldwide has become a multimodal-multimedia one on a mobile smartphone. Multimodal-multimedia interaction and communication technologies have been deployed widely in a variety of consumer products, including smartwatches and smartphones, video conferencing and collaborative systems, augmented and virtual reality systems, smart cars, and elsewhere. In addition, multimedia search and recommendation engines have been developed by major IT companies and social media platforms.

With these transformative technologies and the rapidly changing global R&D landscape, the MM community is now faced with many new opportunities and uncertainties. With the presence of open source dissemination platforms, pervasive computing resources, and advances in deep neural networks, new research results are being discovered at an unprecedented pace. In addition, the rapid exchange and influence of ideas across traditional discipline boundaries have made the emphasis on multimedia multimodal research even more important than before. To seize these opportunities and respond to the challenges, we organized a workshop to specifically address and brainstorm the challenges, opportunities, and research roadmaps for MM research.

The two-day workshop, held on March 30 and 31, 2017 in Washington DC, was sponsored by the Information and Intelligent Systems Division of the National Science Foundation of the

United States. Twenty-three (23) invited participants were asked to review and identify core research areas and application domains of importance to the MM field over the next 10-15 year timeframe. Important topics were selected through discussion and consensus, and then discussed in depth in breakout groups. Breakout groups reported initial discussion results to the whole group, who continued with further extensive deliberation. For each identified topic, a summary was produced after the workshop to describe the main findings, including the state of the art, challenges, and research roadmaps planned for the next 5, 10, and 15 years in the identified area.

Major Research Areas in Multimodal and Multimedia

Given the breadth and depth of the MM field, the workshop participants identified a set of core research areas, whose advances can be integrated to build transformative applications with high impact on industries and society. We first summarize each of the core areas below, followed by discussion of several major application domains. Detailed descriptions of the state of the art, challenges, and anticipated milestones in each area can be found in the full workshop report [1]. A set of nine key research challenges that arise frequently across multiple research areas are further identified and summarized in Section 2 of this paper. The following figure shows an overview of our approach to compile these key research challenges.

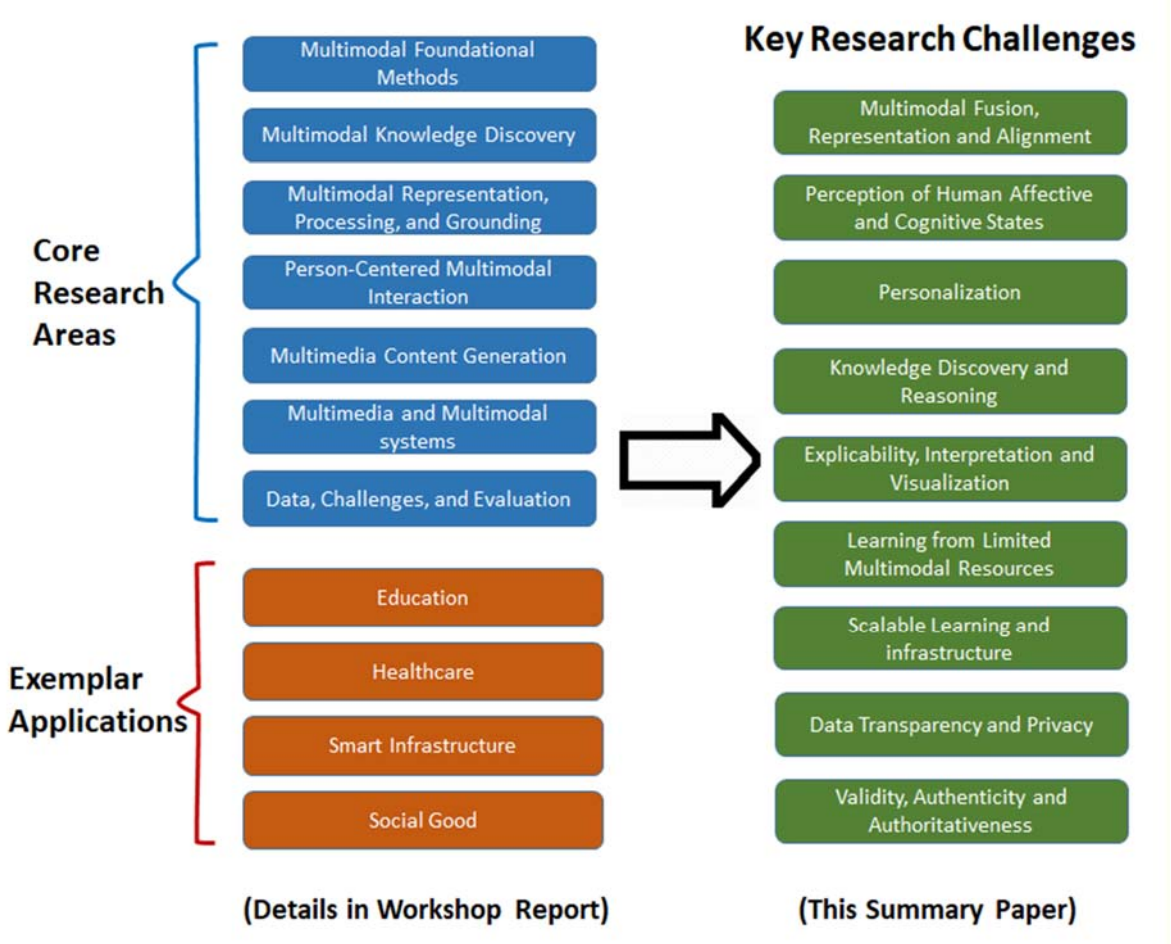

Figure 1: Research Roadmap: Research Areas, Applications and Challenges

We summarize below 7 core research areas for multimodal and multimedia research taken from the full workshop report [1]:

- **Multimodal Foundational Methods**: An important research endeavor is to develop foundational theory and methods to analyze, model, predict and represent multimodal information. Theory is required to guide optimal multimodal system complementarity to achieve ultra-reliable performance levels without relying on trial and error. The methods are characterized as foundational since they are relevant to many applications and research topics. We can group these methods into five classes: representation, alignment, fusion, translation and co-learning.
- **Multimodal Knowledge Discovery**: Knowledge discovery is a fundamental scientific process that historically has aimed to establish cause and effect relations among variables. Since multimodal-multimedia data can be analyzed at multiple levels (signal, activity pattern, representational, transactional, etc.), a vast multi-dimensional space is created for knowledge discovery. Multimodal-multimedia data brings new challenges given its richness, density, contextuality and temporal components. Different strategies can be explored for multimedia knowledge discovery: cross-media discovery studying the intersection of modalities, multi-level multimodal integration, cross-cultural knowledge and unsupervised methods.
- **Multimodal Representation, Processing and Grounding**: Grounding is the connection of lingual expressions with the continuous reality in all its complexity, including other modalities such as visual and acoustic. Studying grounding has the potential to create more natural and at the same time richer and deeper representations of scenes, actions, and events. These findings will benefit communication with autonomous systems (e.g., human-robot communication) as well as knowledge discovery and construction (e.g., common sense knowledge or knowledge population from live sources).
- **Person-Centered Multimodal Interaction**: The unprecedented amounts of online multimodal information available from millions of people enable a better understanding of individuals and their unique characteristics, such as personality, motivations, and interests, at scale. In turn, such understanding of individuals enables person-centered, multimodal interaction between human beings as well as between an individual and an Artificial Intelligence (AI) agent. This research area include three broad sub-areas: multimodal human behavior analysis, multimodal computational psychology and Person-centered multimodal persuasion.
- **Multimedia Content Generation**: The goal of content generation can be defined as a customized, personalized and tractable user experience for simulation, synthesis or entertainment purposes. This can apply to situations where content is generated on-demand for a particular user or groups of users, based on merging a flexible content model with a user profile. Content generation requires applying personal or situational preferences to a content model in order to generate a coherent (multimedia) experience.
- **Multimedia and Multimodal Systems**: Multimedia systems research lies between multimedia applications and the underlying systems that either compute or deliver the expected result. Multimedia systems research focuses on being able to provide (i) mechanisms to run the application in real-time or near real-time, (ii) mechanisms to allow

the application to scale to large numbers, or (iii) provide the best quality-of-experience to the user.

- **Data Challenges and Evaluation**: Useful data is information that provides value possibly through a variety of ways, including training of algorithms, testing, or data used in an operational setting. Scientific data must have mechanisms to enable replication, maintain redundancy, and enable reuse. The idea brings to the forefront challenges in metadata definition, alignment, error in values, holes, semantics of the value, and purpose of acquisition. New techniques are needed to support the acquisition of high-quality and relevant datasets. Development of automated and streamlined data pre-processing methods also are needed to reduce the time and cost of mining multimodal-multimedia data.

Applications of Multimodal and Multimedia Research

In addition to the core research areas listed above, participants identified a few key applications, synergistic with the National Academy of Engineering grand challenges, to exemplify important domains that call for significant advances across multiple MM fundamental areas in the next decade.

- **Education**: Advances in multimodal-multimedia technology have the potential to improve educational outcomes by substantially stimulating learning, and to assess the outcomes more accurately. Multimodal learning analytics are modeling richer data sources like speech, writing, images, and their combination. These methods are beginning to predict learners' actual motivation, cognitive state, and their development of expertise.
- **Healthcare**: Advances in multimedia analytics for healthcare can positively impact areas of general healthcare applications, and personal health, as a result of applications that can analyze personal health data and other factors such as rapid and continued growth in patient-health records. The challenges include clinical protocols on smaller datasets, opportunity for population scale research and emergence of personalized wearable sensors and smart devices.
- **Smart Infrastructure**: Smart infrastructure represents the deep embedding of sensing, computing, communication, decision making and actuation into the traditional physical and human infrastructures for the purpose of increasing efficiency, resiliency and safety. It requires the understanding of embedding multi-sensory devices, the collection and correlation of multimedia data and metadata, utilization of context, multi-modal processing and analytics issues, multi-modal communication and time-sensitive streaming problems, dynamic actuation and adaptation according to human and physical constraints.
- **Social Good**: The aim is to outline a vision, research and directions necessary for multimedia technologies to enhance our communities and human society in general. While multimodal research presents many challenges of genre collection and genre integration, it also presents major promise for tackling critical issues affecting the welfare of a society. Areas such as transportation, home care, security, and journalism are all areas in which multimodal efforts can help improve our society.

2. Key Research Challenges and Roadmaps

For each identified core research area and application, the workshop participants reviewed the state of the art, discussed research challenges, and identified roadmaps and milestones required to address the open challenges. Full reports of these core research areas are included in corresponding chapters of the full workshop report [1]. In our analysis of the report, we further identified a set of nine key challenges that appear frequently across research areas and applications, which can be used to guide the research direction of the MM research community in the next decade. We summarize these nine key research challenges and the milestones and roadmaps required to address these key challenges below. Note the numerical order does not represent any prioritization of the challenges.

Key Challenges	**Milestones and Roadmap**
Challenge 1: Multimodal Theory, Fusion, Representation and Alignment	● new neural representation learning approaches which can handle a large number of modalities and information views ● new temporal representation and alignment models ● new multimodal fusion techniques that successfully generalize to multiple tasks (N>100)
Challenge 2: Perception of Human Affective and Cognitive States	● algorithms able to interpret affective and cognitive states in small group or even largely populated environments ● recognition algorithms that take into account cultural influences and differences ● analysis algorithms robust and extensible in analyzing human affective and cognitive states across different tasks and contexts
Challenge 3: Personalization	● personalization of multimedia analytic and content generation models in the presence of *limited and possibly insufficient* data ● personalization from *real-time and dynamic flow* of multimodal information from the user ● providing individuals control over how their personal data and system representations are used and shared
Challenge 4: Knowledge Discovery and Reasoning	● methodologies to conduct knowledge extraction and representation within *dynamic and large-scale datasets*

	● methods to automate testing the robustness and generality of scientific findings ● continuous proactive knowledge discovery and prediction from streaming data, with ultra-reliable accuracy based on strategic fusion of data sources
Challenge 5: Explicability, Interpretation and Visualization	● interpretable models that involve two or more modalities ● interpretable representations and learning methods for complex knowledge involving complex events and relations ● Incorporation of bio-inspired design strategies into neural network machine learning models
Challenge 6: Learning from Limited Multimodal Resources	● new methods that can learn multimodal interaction from unlabeled data or just a limited amount of labeled multimodal data ● multimodal alignment and fusion techniques that facilitate effective use of existing multimodal data sets in real-world application domains ● crowdsourcing approaches to creating high-quality multimodal data resources ● multidisciplinary collaboration to obtain datasets from primary domains (e.g., medical)
Challenge 7: Scalable Learning and infrastructure	● new tools and testbeds for collecting, fusing, analyzing, and sharing multimodal data from heterogeneous sensors ● next-generation multimedia system framework that can bridge content understanding with multimedia system-level support ● integrated and compatible multi-modal technologies for diverse application sectors

	● evolvable and adaptable multimodal multi-sensor smart infrastructures that can sustain over a long term
Challenge 8: Data Transparency, Fairness and Privacy	● understand inherent bias and skew in multimedia data; building transparency, fairness and the ability to interrogate multimodal algorithms, inferences and systems. ● balance when and how to acquire personal information to establish knowledge bases, while appropriately preserving privacy and data ownership ● develop effective privacy controls for personal multimedia data ● develop secure data storage and analysis protocols ● develop international policy and standards for rich personal data
Challenge 9: Validity, Authenticity and Authoritativeness	● tools for multimedia forensics that allow identification of media tampering and allow full examination of the original context by bringing together all available sources, text, audio and multimedia evidence ● automatic and indelible determination of provenance to establish authenticity of documents, fragments or collections ● automated fact-checking systems to provide context together with confidence ratings of authoritativeness, while being able to identify explicitly false data, biased data, and propaganda that is predatory rather than prosocial

Table 1. Key Research Challenges and Milestones for MM Research in the Next 10 years

Challenge 1: Multimodal Theory, Fusion, Representation and Alignment
Central to many multimodal and multimedia research endeavors is the challenge of representing, aligning and fusing information for multiple modalities. This information is often represented heterogeneously such as the symbols representing spoken words and the pixels representing visual images. Multimodal *representations* are designed to learn how to summarize and represent multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. Multimodal *alignment* is to identify the direct links between sub-elements from two or more different modalities. The natural asynchrony between spoken words and gestures is a good example where alignment is needed. Multimodal *fusion* joins information from two or more modalities to perform a prediction, such audio-visual speech recognition. Addressing these challenges of representation, fusion and alignment are fundamental if we want to be able to discover, extract and summarize knowledge from very large and heterogeneous data sources such as found on the internet (e.g., Wikipedia) [See Chapter 3 of the full report]. This is also important for recognizing multimodal person-centric behaviors such as emotions and social signals which are expressed through language, visual and acoustic modalities [See Chapter 5].

The roadmap for this challenge requires (1) to develop new representation learning approaches which can handle a large number of modalities and information views, going beyond the current state-of-the-art which usually studies two modalities; new neural approaches for multimodal representation to understand the geometry of the multi-modal semantic space (2) to create new temporal representation and alignment models which correctly segment relevant sub-events, model their long-range dependencies and summarize the essential content of long videos or sequential data; and (3) to create new multimodal fusion techniques that successfully generalize to multiple tasks (N>100) by taking advantage of the complementarity of these prediction tasks (4) formulate new theory to guide optimal multimodal system complementarity in order to achieve ultra-reliable performance levels without relying on trial and error.

Challenge 2: Perception of Human Affective and Cognitive States
A human-centric challenge of multimodal research is to automatically recognize, model and even generate emotions and other affective and cognitive states naturally expressed during social interactions. Affective states can include sentiments, attitudes or personality traits. These are complemented with cognitive states which can include attention, engagement, curiosity, domain knowledge, deceptive intent, and many others. Affective and cognitive states are important for many multimodal applications such as education where perceiving the levels of curiosity and engagement of a student are often essential to successful learning [Chapter 9]. Recognizing human affective and cognitive state requires the integration from multiple modalities, including language, acoustic and visual modalities [Chapter 2]. In other words, what you say, how you say it and the gesture accompanying it. These states can be expressed when interacting with a computer (e.g., interaction with a conversational assistant such as Alexa, Siri, Google Assistant or Cortana), or when observing natural human social interactions (e.g., doctor-patient interactions). It requires to contextualize the interpretation of the multimodal behaviors with the previous interactions as well as the environment where they happen.

The roadmap for this challenge requires (1) to create algorithms able to interpret affective and cognitive states in small group or even largely populated environments, where more than two people are interacting with each other; (2) to develop recognition algorithms that take into account cultural influences and differences to accurately when interpreting human multimodal behavior and characteristics across the globe; and (3) to create analysis algorithms robust and extensible in analyzing human affective and cognitive states across different tasks and contexts, such as healthcare, education and personal robotics applications, and (4) creating bias-free interpretations of human affect and cognition.

Challenge 3: Personalization

Modeling of human multimedia content and behaviors will require new technologies that are able to adapt to user's idiosyncrasies and learn their respective preferences. This is motivated by the fact that individual's desires and intents are different based on their own life experiences, their current needs and the context of the interaction. For example, when visiting a city like Paris, some people may prefer a more culture-oriented experience while others would desire an adventurous experience. This personalization requirement is particularly relevant to multimedia content generation where content is generated on-demand for a specific user (or groups of users) based on merging a flexible content model with a user profile [chapter 6]. The goal of such content generation is to provide a customized, personalized and tractable user experience for simulation, synthesis or entertainment purposes. Personalization is also key for person-centered multimodal interaction where interpretation of user's multimodal behaviors should be taking into consideration the interaction context and as importantly the preferences, desires and idiosyncrasies of the user [chapter 5]. Personalisation is challenging. This is because much of the behavioral data is heavy tailed—most data is generated by a few people, making it hard to make predictions for the many for whom we have little data. Personalization of human multimodal behaviors and content is also central to two applications, namely education and healthcare [chapters 9 and 10]. In education, personalizing the learning applications is based on the multimodal assessment of students' cognitive state and actual learning achievements. In healthcare, personalized medicine where treatment is tailored to the individual characteristics of each patient requires understanding information from multiple modalities and sensors.

The roadmap for this challenge requires (1) to learn personalization of multimedia analytic and content generation models in the presence of *limited and possibly insufficient* data from the specific user or group of users, (2) to perform personalization from *real-time and dynamic flow* of multimodal information from the user, requiring the multimedia computational framework to continuously monitor and adapt to behavioral changes, and (3) to allow control over how the personal representations and data are used and shared, including the ability to recall or withdraw their content's use on demand.

Challenge 4: Knowledge Discovery and Reasoning

The richness of multimodal and multimedia data has the potential to change the way we discover and reason about knowledge. As an illustration of their richness, multimedia signals typically involve multiple time-synchronized data streams. Information within the multimedia data streams can be analyzed across multiple levels, including signal, activity pattern, representational, and transactional [chapter 3]. Knowledge discovery is a fundamental scientific

process to collect and evaluate data in a manner that pursues a process of elimination, or that isolates individual variables, in order to assess their causal impact. This process has centered on hypothesis testing, and the establishment of scientific methods that can support accurate inferences about cause and effect relations among the variables of interest. In order to understand the full scope and importance of a new finding or theory and its applicability, the scientific process also pursues a lengthy process of replicating and testing generality under varied conditions. With an eye toward artificial intelligence and machine learning, there is a need for developing algorithms for extracting, representing and modeling knowledge from vast amounts of data [chapter 8]. Reasoning on these multimedia content sources requires grounding and relationships between heterogeneous modalities, linking abstract concepts such as linguistic elements to real-world physical object, people and images [chapter 4].

The roadmap for this challenge requires (1) to develop methodologies to conduct knowledge extraction and representation within *dynamic and large-scale datasets*, going beyond static knowledge models, (2) methods to automate testing the robustness and generality of scientific findings; (3) continuous proactive knowledge discovery and prediction from streaming data, with ultra-reliable accuracy based on strategic fusion of data sources;

Challenge 5: Explicability, Interpretation and Visualization

As machine learning becomes prevalent in multimodal research, a key open challenge is how to enhance the interpretability and visualization of the multimodal models [Chap. 2, 3, 8]. We should design our multimodal models with the goal of being able to interpret the intermediate representations and alignments between modalities. Today, we still lack robust ways for visualizing the interaction among multimodal content, and for understanding how multimedia knowledge (including entities, events, and relations) are captured in related multimodal representations or ideally a single common representation. This is particularly important when we start using these multimodal models in real-world scenarios such as health care and law enforcement where users will need to trust and understand, at least at a high level, what the computer is doing. Visualization of intermediate representations in deep learning model also offers opportunities to understand the roles and interdependency of various layers or modules in the architecture. This provides opportunities to discover new insights that can be used to further improve the design of the architecture for better efficacy or efficiency. Finally, visualization and interpretation methods will facilitate comparisons with discoveries in biological neural science and incorporate such bio-inspired approaches to the design of new neural network models. Milestones towards addressing the challenges include (1) extending current progresses in interpretable models for single modalities (language or vision) to cases that involve two or more modalities, (2) interpretable representations and learning methods for complex knowledge involving complex events and relations, and (3) success in incorporating bio-inspired design strategies into neural network machine learning models, and (4) extended dialogue-based query capabilities for obtaining specific relevant explanations” (i.e., in addition to visualization, which is not enough alone.

Challenge 6: Learning from Limited Multimodal Resources

Despite the rapid progress in machine learning, scarcity of labeled data continues to present major challenges for developing robust machine learning models. In response to this, various

research has been proposed to tackle resource-limited learning problems such as unsupervised or semi-supervised learning, weakly supervised learning, transfer learning, and zero- or few-shot learning. This problem becomes even more acute for tasks involving multiple modalities. For example, for many fundamental tasks such as multimodal alignment or translation (Chap. 2), annotating alignment relations or acquiring translated data among multiple modalities is much more difficult than the cases of single modality. As another example, for multimodal grounding (Chap. 4) that aims at linking linguistic mentions to specific sensory data instances, datasets capturing realistic grounding relations under natural settings (as opposed to simulated contexts usually used in crowdsourcing based annotation) are quite limited. Finally, in many application domains such as Healthcare (Chap. 10), there are often data errors or gaps in subsets of modalities that are difficult to discover but have detrimental effect on the final outcomes.

The roadmap for solving these challenges requires (1) new meta-learning methods for understanding multimodal interaction from unlabeled data or just a limited amount of labeled multimodal data; (2) multimodal alignment and fusion techniques that facilitate effective use of existing multimodal data sets in real-world application domains such as healthcare; (3) techniques for improving crowdsourcing approaches to creating high-quality multimodal data resources, such as active learning, incentive creation through “gamification”, and methods to enhance the naturalness of multimodal interaction; (4) multidisciplinary collaboration to obtain datasets from primary domains (e.g., medical).

Challenge 7: Scalable Learning and infrastructure

The impact of multimedia/multimodal technologies will be limited if new theories and tools are not integrated with practical systems equipped with real-time sensors, computing/communication technologies, and real-world applications interacting with users. In the area of knowledge discovery (Chap. 3), there is critical need of large-scale heterogeneous multimedia data collection, alignment, and analysis in order to discover and validate new knowledge from specific application domains. In developing next-generation multimedia systems (Chap. 7), it’s recognized more efforts should be made to exploit the synergy between intelligent content understanding and multimedia system design. For example, intelligent content understanding produces semantic information that’s valuable for developing intelligent compression, transmission, or distributed control solutions. In developing future smart infrastructures (Chap. 11), major challenges remain in deep embedding of sensing, computing, communication, decision making and actuation into the traditional physical and human infrastructures that can increase efficiency, resiliency and safety, and can also sustain, evolve, and adapt over a long time.

To address these challenges, advances are called for a few milestones: (1) new tools and testbeds for collecting, fusing, analyzing, and sharing multimodal data from heterogeneous sensors, (2) next-generation multimedia system framework that can bridge content understanding with multimedia system-level support (e.g., system control for Quality of Experience management), (3) integrated and compatible multi-modal technologies for diverse application sectors (e.g., rail, automobiles, ships for smart transportation), and finally (4)

evolvable and adaptable multimodal multi-sensor smart infrastructures that can sustain over a long term.

Challenge 8: Data Transparency, Fairness and Privacy
Multimodal data drives much current innovation and is a source of great profits. Many users are providing data without understanding the privacy implications. Due to sparsity of data, much data collection for research and commercial systems uses any available data for learning, without much consideration of bias, skew or subtle domain mismatch [chapter 5]. In addition, fairness in the inference process is important: complex inference algorithms can vary with attributes of gender and race. This variation can run afoul of the law (e.g. recommendations in housing related search). In addition, there is no transparent or explainable representation in the most accurate deep learning methods [chapter 4]. The most affected applications are user-centric, enabling hyper-individualized user interfaces and interactions. Specifically, education [chapter 9] and healthcare [chapter 10] are areas where personal data is critical. Future education will exploit individualized learning strategies, requiring data suitable for the individual, but educational privacy controls restrict sharing of data. Healthcare's big promise is in personalized treatments, yet privacy restrictions on health records preclude sharing useful data.

The roadmap for this requires solution for these top challenges here: (1) Developing approaches to understanding inherent bias and skew in multimedia data and building interrogability, explainability and fairness into multimodal learning systems so that reasons for improper results can be identified and mitigated. (2) Healthcare research will need to balance exploiting personal variabilities in treatment by expanding and amplifying our existing biomedical knowledge base, while also addressing issues of privacy, data ownership and attribution of provenance. (3) Development of effective privacy controls for personal multimedia data. This includes user interfaces for easy specification of differential privacy constraints, revocability of permissions, and universally implemented protocols that grant individuals the ability to understand and modify consent. Essentially, this needs to re-balance the information asymmetry between the users and data collectors. (4) Development of international policy and standards for rich personal data to support future applications and services in a responsible manner for broad populations.

Challenge 9: Validity, Authenticity and Authoritativeness
As a tsunami of multimedia content is constantly generated both manually and automatically, the problem of information provenance takes on increased urgency. Understanding the heritage of information, revealing the underlying narratives used to structure content, verifying the authenticity of individual content sequences and explicitly documenting the full context of capture and use becomes essential in maintaining an open information society. There is a strong need to develop methods that can provide users with confidence that the content of any message embedded in the media is verifiable and accurate. To date, virtually nothing protects the rights of content consumers in the same way that the rights of users of consumer products are protected. In an information society, understanding the basic qualities of that information is essential. Among the application areas and societal institutions that would be most heavily impacted are education (Chapter 9), journalism, advertising, and politics (Chapter 12), as well as the foundations of multimedia analysis (Chapter 2) and knowledge discovery (Chapter 3).

Without verifiably accurate data, knowledge discovery cannot occur, and democratic societies will disintegrate.

The roadmap for verification, authentication of provenance, and assessment of authoritativeness includes the following key challenges: 1) Tools to perform increasingly sophisticated multimedia forensics that allow detection of media that has been tampered with and allow full examination of the original context by bringing together all available sources, text, audio and multimedia evidence. 2) Automatic and indelible determination of provenance to establish authenticity of documents, fragments or collections as a whole. Beyond tools for determining the first post of a piece of information and the authenticity of a location in an image, an ‘objective’ multimedia documentation of what has occurred including different points of view should be available for most events either as explicitly recorded or retrospectively reconstructed. 3) Fully automated fact-checking systems to provide context and background together with confidence ratings in their authoritativeness. The systems must be capable of identifying explicitly false data, biased data, and propaganda that is predatory rather than prosocial in nature. Automated intelligent systems will need to make sense of the digital maelstrom and allow users to identify the major voices amongst the crowds.

3. Conclusions

We hope by identifying the core research areas and applications, understanding the state of the art and needed advances in in each area, and summarizing the key research challenges cross cutting multiple areas and applications will provide a useful roadmap guiding the research effort of our community in the next decade. We see this workshop and report as a first step to initiate this important discussion around multimodal and multimedia research roadmap. We anticipate new areas and challenges to emerge as the rapid progresses in the multimedia and multimodal field continue. Additional workshops and efforts will certainly be needed in the future to address the new challenges and opportunities and augment the current report with the additional findings.

References:

[1] Shih-Fu Chang, Alex Hauptmann, Louis-Philippe Morency, Sameer Antani, Dick Bulterman, Carlos Busso, Joyce Chai, Julia Hirschberg, Ramesh Jain, Ketan Mayer-Patel, Reuven Meth, Raymond Mooney, Klara Nahrstedt, Shri Narayanan, Prem Natarajan, Sharon Oviatt, Balakrishnan Prabhakaran, Arnold Smeulders, Hari Sundaram, Zhengyou Zhang, Michelle Zhou, *"Report of 2017 NSF Workshop on Multimedia Challenges, Opportunities and Research Directions," arXiv preprint arXiv:1908.02308* (2019).