SCADS 2024 Problem Book (S1).

Violet B., Susanna Bitters, John M (S2). Conroy, R (S3). Jordan Crouser, Sue Mi Kim, J (S4). Bowman Light, Neil P (S5). Molino, Amanda Peterson, Elizabeth Richerson, Stephen Shauger, Ben Strickson, Aaron W., Julia S (S6). Yang (S7).

April 2024 (S8).

### 1 Introduction (S9).

The Summer Conference on Applied Data Science (SCADS) is an annual eight-week summer research program held at the Laboratory for Analytic Sciences (LAS) that is focused on the research and development of Artificial Intelligence (AI) to assist US Intelligence Community (IC) data analysts (S10). SCADS identifies a multi-year Grand Challenge to focus the research toward solving a current mission challenge (S11). SCADS brings together approximately fifty personnel each year, including faculty, graduate students, industry professionals, researchers at National Laboratories and Federally Funded Research and Development Centers (FFRDCs), and government employees from the IC, to advance the research on the current Grand Challenge. (S12).

Grand Challenge: Generate tailored daily reports for knowledge workers that capture information relevant to their individual objectives and interests (S13).

The SCADS grand challenge is intended to be a multi-year unifying research goal of creating Tailored Daily Reports (TLDRs) for individual knowledge workers within the IC (S14). These reports would be similar in some sense to the well-known President’s Daily Brief or to established commercial news aggregators, but would include a mix of classified and unclassified material, combine information from a variety of modes and formats, and be tailored to the interests and responsibilities of the individual IC worker (S15).

In 2024, we continue working toward creating TLDRs and directed research efforts will fall under the broad areas of Automatic Summarization, Recommendation Systems, and Human-Machine Interaction (S16). Each of these areas will also include a special emphasis on explainability and contextualization, as well as knowledge representation and dataset creation and curation (S17). Efforts will also include development projects that are aimed at bringing together the different areas into a demonstration application to model how the individual research projects come together to support a TLDR (S18).

In this document we present relevant critical challenges and research questions that address various aspects of the grand challenge, briefly discuss the motivation for those challenges and questions, and provide references and other resources for further exploration (S19). We have organized this document by focus areas, and note that some challenges and questions presented might span multiple focus areas (S20). For each critical challenge and research question, we attempt to identify the scope (short, medium, or long-term) of the challenge or question, and also which other challenges and questions are related (S21).

Throughout this document we attempt to identify research areas which may be served by use of Large Language Models (LLMs) which have been an area of great focus in the field recently (S22). Two ubiquitous LLMs are GPT-3.5 and GPT-4, versions of which are incorporated into ChatGPT (S23).

### 2 Automatic Summarization (S24).

Key to the creation of a TLDR will be the automatic summarization of large corpora of documents (S25). This process will involve both extractive and abstractive summarization (S26). While summarization for SCADS is primarily concerned with text, research in multi-modal summarization techniques is also encouraged (S27). Researchers would have access to existing summarization engines and be able to improve and augment them (S28). Traditionally, automatic summarization has fallen into two categories: extractive and abstractive (S29).

Extractive summarization is the process of extracting key sub-portions of a document, typically at the sentence level, and compiling those sub-portions into a readable and salient summary of the document (S30). The main extractive summarization package we will use at SCADS is occams [1] (S31). The occams package employs an optimization-based extractive algorithm (S32). This family of methods is known to perform best among extractive systems on a wide range of evaluation data (S33). For a survey of the summarization problem and classical extractive summarization methods, we recommend [2] (S34). Also, see [3] for an overview with a description of some successful neural-based abstractive summarizers (S35).

Abstractive summarization is the process of generating novel prose providing a readable and salient summary of a document based on its content, rather than extracting sub-portions of the document as in extractive summarization (S36). Large Language Models (LLMs) employ abstractive summarization (S37). For SCADS, we will focus on models made available by Hugging Face, the OpenAI API, and other models as applicable (S38). Open-source systems available via Hugging Face are recent small to midsize neural systems using variations of the transformer language model, the first of which was BERT [4] (S39). Models of this size serve as practical examples of abstractive summarization capabilities in a computing environment similar to what might be expected for a TLDR prototype or operational system (S40). Other larger language models, such as GPT-3 and -4, are also eligible for inclusion during SCADS (S41). The focus on such large language models might shift to address how such models might be practically incorporated into a TLDR scenario, for example, through exploration of hybrid implementations with an extractive summarizer, careful prompt creation, or Retrieval Augmented Generation (RAG) [5] (S42).

For all automatically created summaries, it will be important to provide context for where the information contained in the summary originated (S43). This might be in the form of references, citations, or other methods enabling users to move from the summary to the underlying source data (S44). See 4.4 for more discussion of providing context for content included in a TLDR (S45).

Critical Challenge 2.A Develop summaries of documents personalized to a user’s interests, employing abstractive extractive, and/or hybrid methods (S46).

As improvements are made in the language model space, the scope of summarization is increasing (S47). Topics in question answering, information retrieval, and entity extraction are becoming more mainstream for summarization engines within traditional summarization, there are various methods to investigate that focus on different aspects of the text (S48). A past Document Understanding Conference hosted by NIST investigated topic-based summarization (S49). Other potential areas could be guided and aspect summaries, all of which incorporate named entities. (S50).