The rise of cold data. Driven by the promise of machine learning and data analytics, enterprises routinely gather vast amounts of data from diverse data sources. Analysts have reported that enterprise data is growing 40% annually and will account for 60% of the 160 Zettabyte Global Datasphere by 2025. However, not all data is accessed uniformly. Studies have reported that only 20% of data stored is performance-critical and accessed frequently. The remaining 80% is cold and accessed infrequently. Historic data used for trend forecasting, archival data stored for meeting legal and regulatory audits, and backup data accessed during failures, are examples of such cold data. Cold data has been identified as the fastest growing data segment with a 60% annual growth rate, and also as the segment with the longest lifetime (window between creation and deletion date) with retention periods lasting 50–60 years. Thus, enterprises are in desperate need of cost-effective options for long-term storage of cold data.

Limitations of current storage media. Unfortunately, all current storage media suffer from two fundamental limitations that make them unsuitable for storing cold data. First, all current media suffer from density scaling limitations resulting in storage capacity improving at a much slower rate than the rate of data growth. For instance, Hard Disk Drive (HDD) and magnetic tape capacity is improving only 16-33% annually, which is much lower than the 60% growth rate of cold data. Second, all current media have very limited lifetime and hence, suffer from the media obsolescence problem. For instance, HDD and tape have a lifetime of 5–20 years. Enterprises regularly archive data for much longer time frames. Thus, using HDD or tape for archival leads to constant data migration with each new generation to deal with device failures and technology upgrades. A recent article summarized the financial impact of such media obsolescence on the movie industry. Due to these limitations, there has been an industry-wide consensus that no storage media available today will be unable to meet the Zettabyte-scale cold data storage needs of the future.

This project details a research plan to build storage technology that can overcome the aforementioned limitations using a radically new media, namely Synthetic DNA. Deoxyribo Nucleic Acid (DNA), is a macro-molecule that is composed of smaller molecules called nucleotides. There are four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA used for data storage is a single-stranded sequence of these nucleotides, also referred to as an oligonucleotide (oligo). Using DNA as a digital storage medium requires mapping digital data onto a sequence of As,Cs,Gs, and Ts, using an encoding algorithm. Once encoded, the nucleotide sequence is used to synthesize DNA using a chemical process that assembles the DNA one nucleotide at a time. Data stored in DNA is read back by sequencing the DNA molecules and decoding the information back to the original digital data.

DNA opportunities. DNA possesses three key advantages over current storage media. First, it is an extremely dense three-dimensional storage medium with a capacity of storing 1 Exabyte/mm3 which is eight orders of magnitude higher than magnetic tape–the densest medium available today. Second, DNA is very durable and can last millennia in a cold, dry, dark environment. A recent project that attempted to resurrect the Woolly Mammoth using DNA extracted from permafrost fossils that are 5000 years old is testament to the durability of DNA even under adverse conditions. Third, the density of DNA is biologically

fixed. Thus, data once stored in DNA can be left untouched without repeated migration to deal with technology upgrades. As a result, DNA does not suffer from media obsolescence making it an ideal candidate as a media for storing cold data. Due to these benefits, DNA has received much attention from the industry recently. The US government has announced the MIST program to accelerate innovation in DNA storage. Microsoft has announced plans to make experimental DNA storage available in the Azure cloud by 2020. However, current work on DNA data storage is only in its nascency, and has primarily focused on encoding techniques for storing instructured data like photos and videos.

Goal and objective. Our goal is to open up DNA for wider adoption by using it as storage media for future enterprise data analytics platforms. Modern analytics platforms have two salient properties. First, data stored in these platforms is structured with a well-defined data model. Second, applications built on these platforms have varying precision requirements. The objective of this project is to exploit these properties and build a radically new storage stack for future data analytics platforms that uses DNA, instead of tape or HDD, for cold data storage. This objective creates several challenges that require fundamentally rethinking all major aspects of data management, from data layout to data processing. We overcome these challenges by adopting an inter-disciplinary research agenda that spans data management, information theory, computational biology, algorithmics, and high-performance computing (HPC).