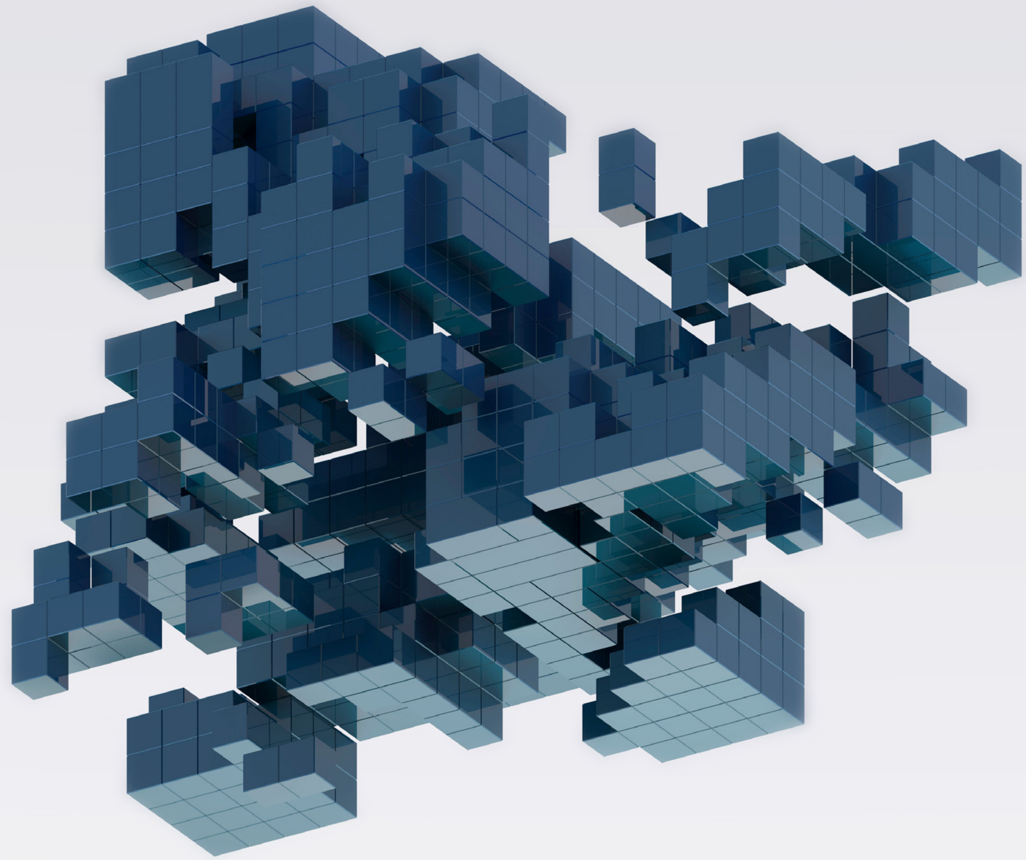


Future Directions in Compressed Sensing and the Integration of Sensing and Processing

What Can and Should We Know by 2030?

Robert Calderbank, *Duke University*
Guillermo Sapiro, *Duke University*

Prepared by Brian Hider and Jeremy Zeigler
Virginia Tech Applied Research Corporation

Workshop funded by the Basic Research Office, Office of the Assistant Secretary of Defense for Research & Engineering. This report does not necessarily reflect the policies or positions of the US Department of Defense

VT-ARC
Virginia Tech
Applied Research Corporation

> Innovation is the key to the future, but basic research is the key to future innovation.
>
> *–Jerome Isaac Friedman, Nobel Prize Recipient (1990)*

Preface

Over the past century, science and technology has brought remarkable new capabilities to all sectors of the economy; from telecommunications, energy, and electronics to medicine, transportation and defense. Technologies that were fantasy decades ago, such as the internet and mobile devices, now inform the way we live, work, and interact with our environment. Key to this technological progress is the capacity of the global basic research community to create new knowledge and to develop new insights in science, technology, and engineering. Understanding the trajectories of this fundamental research, within the context of global challenges, empowers stakeholders to identify and seize potential opportunities.

The Future Directions Workshop series, sponsored by the Basic Research Office of the Office of the Assistant Secretary of Defense for Research and Engineering, seeks to examine emerging research and engineering areas that are most likely to transform future technology capabilities.

These workshops gather distinguished academic and industry researchers from the world's top research institutions to engage in an interactive dialogue about the promises and challenges of these emerging basic research areas and how they could impact future capabilities. Chaired by leaders in the field, these workshops encourage unfettered considerations of the prospects of fundamental science areas from the most talented minds in the research community.

Reports from the Future Direction Workshop series capture these discussions and therefore play a vital role in the discussion of basic research priorities. In each report, participants are challenged to address the following important questions:

- How might the research impact science and technology capabilities of the future?
- What is the possible trajectory of scientific achievement over the next 10–15 years?
- What are the most fundamental challenges to progress?

This report is the product of a workshop held January 11–12, 2016 at Duke University in Raleigh-Durham, North Carolina on Compressed Sensing and the Integration of Sensing and Processing. It is intended as a resource to the S&T community including the broader federal funding community, federal laboratories, domestic industrial base, and academia.

Executive Summary

As it becomes possible to access ever larger amounts of data, it becomes increasingly important to develop methods of *smart sensing* that are able to support information-based decisions. A trend common to all disciplines and areas is the integration of sensing and (task-oriented) processing as one unit. At a workshop, held at Duke University on January 11–12, 2016, thirty distinguished researchers from academia and industry gathered to discuss the opportunities and challenges of Compressed Sensing research to intelligently address the ever increasing data produced by modern technology. This report captures those discussions and presents the consensus opinion about the state of the field, the challenges to progress and the trajectory of research for the next 10 and 15 years.

The workshop participants organized the current and future research that are fundamental to progress in sensing and data processing, into four areas: signal acquisition, data models, algorithms and architecture.

They framed important challenges for these research areas within the context of:

- Integration of sensing and data processing
- Integration of distributed and multimodal sensing
- Improved analog-to-digital conversion
- Data models for both large and small datasets
- Algorithms optimized for efficiency and that integrate human-machine interactions
- Performance analytics
- New architecture principles

In addition to discussing important research challenges, the participants discussed the value of targeted investment in infrastructure, including the development of open-source tools and International Centers of Excellence. The participants were generally optimistic about the trajectory of Compressed Sensing research and expect continued research will meet the challenges over the next 10 and 15 years. The specific milestones and time frame for reaching the goals is:

10 year goals

- Sensing+X or Task Oriented Sensing
- Model driven signal acquisition
- New models, with a particular emphasis on the global geometry of data, to support development of non-convex/nonlinear optimization
- Adaptive learning, with a particular emphasis on systems with a human in the loop
- New methods of data dependent regularization
- New interfaces between modeling and computation

15 year goals

- Robust data fusion
- Non-convex optimization
- New methods of predicting performance, with a particular emphasis on deep learning
- Architectural principles for networked data processing systems

Continuing goals

- Model-based simulated data
- Common test frameworks, from data sets to competitions
- Interdisciplinary research programs that encourage theorists and domain scientists to collaborate on building systems

These research goals draw from the disciplines of electrical and computer engineering, statistics, computer science and mathematics. Therefore, an integrated intellectual and capital investment in these disciplines, from the theory to the domain application, is critical for the advancement of knowledge and leadership in *data science*.

It has never been more important to understand first, what information is necessary to operate and compete; and second, how this information is sensed and applied. Investment in basic science is essential to staying ahead of the volume at which data is becoming accessible.

"It has never been more important to understand first, what information is necessary to operate and compete; and second, how this information is sensed and applied."

Introduction

High energy physics depends more and more on massive data sources such as the Large Hadron Collider and Large Synoptic Survey Telescope. The latter is about to transform our understanding of the dynamic universe. Personalized medical care depends on detailed individually collected data, very often including *epigenetic* data merged with demographic and environmental data. Environmental research depends on ever more sophisticated observational data coupled with large model-generated datasets. Internet communication can be defined as a graph with billions of nodes. Tracking and predicting the evolution of these kinds of graphs is one of the grand challenges of modern computer science. In fact, the central issue across a broad spectrum of science is specifying what information is needed to predict the evolution of large complex systems, and answering how this information is obtained and applied.

US leadership of the global services economy depends on the ability to analyze data at massive scale, to distill information quickly, and to act on it rapidly with greater insight than an adversary. Moreover, all of this must be done with consistency.

Advances in *data science* can dramatically alter how data is sensed, stored, interpreted and acted upon, and these developments will in turn have fundamental implications for national security and US scientific and economic leadership.

The Large Synoptic Survey Telescope, in construction on the Cerro Pachon Mountain in Chile will generate 30 terabytes of data a night, every night, for ten years.

As massive as the Internet has become, it will be dwarfed by the Internet of Things that is starting to take shape, where all devices are equipped with diverse sensors and communicate. *(image courtesy of wordstream.com)*

Scale and the Data Dilemma

Scale changes the face of society. The introduction of the assembly line enabled Henry Ford to cut car prices and create a mass market. The US mastered scale and became the preeminent manufacturing economy. In similar fashion, **success and failure in the modern era is a function of excellence in information at scale.** The key to future success is to be able to access *data*, to distill information from such data fast, and to act on it fast with greater insight than the competition. Perhaps most importantly, all of this must be done with consistency.

The world of science used to be a place where it was difficult to collect data. *Sensing* was part of the research challenge. The science community developed sophisticated tools for sorting outliers in small data sets. That world is passing—today we are drowning in data streams and we need to develop the computational science necessary to sense smart and to distill information at massive scale.

What constitutes massive scale? The Large Synoptic Survey Telescope, in construction on the Cerro Pachon Mountain in Chile, will generate 30 terabytes of data a night, every night, for ten years. The recently started BRAIN initiatives are producing terabytes of data a day, even for relatively small 1mm cubic brain regions. Internet communication defines a graph with billions of nodes. Tracking and predicting the evolution of such graphs as they evolve is one of the grand challenges of modern computer science. There is currently an estimated 3.8 trillion photographs, 10% of them taken in the last year. Facebook has about 140 billion images with about 300 million new images a day. YouTube contains in the order of 120 million videos and 72 hours of video uploaded every minute. This is of course data, but not necessarily information, and certainly not actionable information.

Numerous additional examples exist in many disciplines. One such example is the Department of Veterans Affairs Office of Research & Development's Millions Veterans Program (MVP). The goal of this national voluntary research program is to study how genes affect health. To do this, MVP will build one of the world's largest medical databases by safely collecting blood samples and health information from one million Veteran volunteers. Data collected from MVP will be stored anonymously for research on diseases like diabetes, cancer, and post-traumatic stress disorder. Google, along with Stanford and Duke University are also collecting large amounts of health data in MVP's Baseline program. Meanwhile, Apple's *Open Source* Research Kit is driving a revolution in global medical research participation.

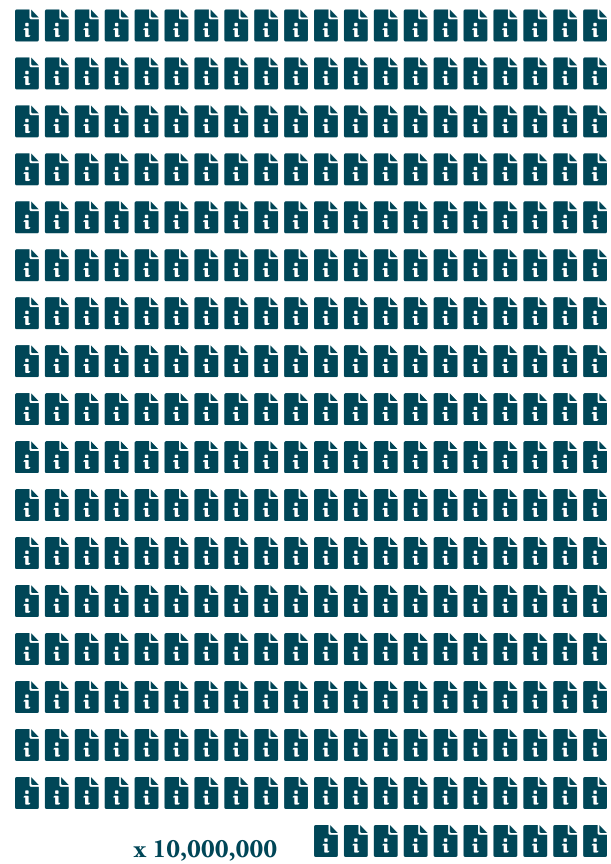
In education, some states such as North Carolina provide pupil information under strict privacy restrictions in order to conduct research toward improving child wellbeing; such data can be combined with medical records as well. Some countries even curate all of their public education data for research purposes. For instance, Plan Ceibal in Uruguay, implementing the "1 to 1" model to introduce Information and Communication Technologies in primary public education. In four years, Plan Ceibal delivered 450,000 laptops to all students and teachers in the primary education system and no-cost Internet access throughout the country, with all the data now being curated for unique child wellbeing studies.

Most data numbers are astonishing. It is expected that by 2020 the amount of digital information in existence will have grown from 3.2 zettabytes today to 40 zettabytes. Every minute we send 204 million emails, generate 1.8 million Facebook likes, and send 278 thousand Tweets. **Google alone processes on average over 40 thousand search queries per second, making it over 3.5 billion in a single day.**

90% of Google's data was created in the last 2 years. To go from data to decision will require excellence in judicious data selection, sensing the environment, marshaling *data*, systems and software for storage and processing, data analysis, and data visualization. These functions form a figurative chain and excellence is required of every link. Once the chain is in place, it is possible to start from a data source and some

Google alone processes on average over 40 thousand search queries per second.

Making it over 3.5 billion in a single day.



application of specific knowledge and to provide insight in seconds, hours, days or weeks rather than months, years or not at all. The chain is of course not necessarily linear/sequential, and not a feed forward system, but an integrated one where all components benefit and influence each other. "Don't sense what you are not going to use" describes how data analysis should feed data acquisition, and the reduction in data itself can help the processing, analysis, and visualization.

We have a new Renaissance paradigm. While we should not expect revolutionary advances to result from an individual Leonardo DaVinci who knows and integrates everything, resulting in revolutionary inventions, we need to have DaVinci-type teams and collaborations, that leverage expertise in all the various *data science* components.

The weakest link in the data science chain is the basic science that will reveal the structure of multimodal data in high dimensional space so that we can make use of that structure in order to make decisions that encompass acquisition to action. Without investment in multidisciplinary basic science and long-term paradigms, we will simply strand all the investments in physical infrastructure that we have made and the US will lose strategic and economic leadership.

It is under this pretext that 30 researchers met at Duke University on January 11th and January 12th to discuss the future direction of *compressive sensing* and the integration of sensing and processing fields. The purpose of the meeting was to explore current and projected issues, topics, and ideas that address the scale and scope of information processing at scale. This report represents the product of these discussions and attempts to portray consensus opinion regarding the salient challenges and trajectories presented during the workshop.

"The weakest link in the data science chain is the basic science that will reveal the structure of multimodal data in high dimensional space so that we can make use of that structure in order to make decisions, from acquisition to action."

The debut of Pope Francis was recorded for posterity by thousands of eyewitnesses packing St. Peter's Square. But eight years earlier, there was nary a cell phone camera in the crowd when Pope Benedict assumed the papal throne, as series of remarkable then-and-now photos reveal. *(image courtesy of nydailynews.com)*

Recent Advances in Data Science and Sensing

The workshop discussed developments in both the applied and basic science areas of *compressive sensing* and the integration of sensing and processing research. This section highlights some of those discussions and offers context to where the future of the field may be headed.

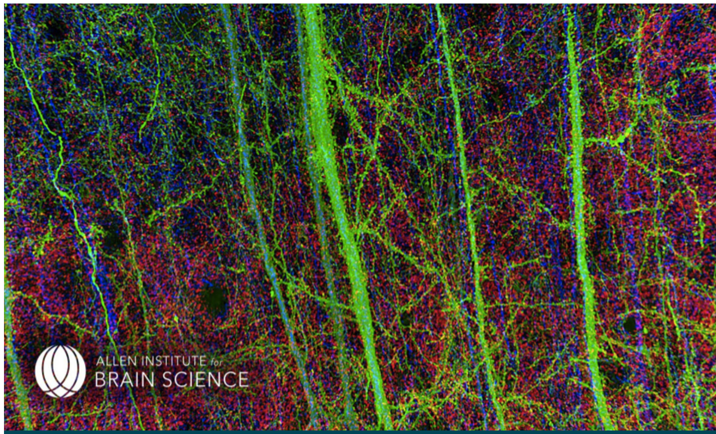
Signal Acquisition

Data acquisition is no longer a fundamental bottleneck. Radar and camera hardware have become physical front ends for massive signal processing problems, leading to algorithms that take full advantage of the physics of acquisition. New developments within image processing include the STORM microscope (Nikon Instruments Inc.), the proof of concept single pixel camera (Rice University), the Coiffman/Golay camera, the DARPA funded gigapixel camera (Duke University), movie acquisition in cryo-tomography, and array tomography for brain data acquisition (Stanford University and the Allen Institute). These are just a few examples, complemented by data simultaneously acquired my millions of users with consumer-grade devices in their pockets.

The recognition that many signals of interest have sparse representations has inspired development of wavelets and *compressed sensing*, leading to a modern sampling theory that is changing the technology landscape. Examples include faster MRI scanners and video cameras. More recent developments include the emergence of nonlinear methods that capture instantaneous frequency and are useful in analyzing almost periodic data such as ECG data.

The sampling budget itself can be reduced by *adaptive sensing*, where the next sample decision depends on the current state of knowledge. Adaptive sensing can be fully automatic, or include a human in the loop, thereby providing a way to scale the impact of limited subject matter expertise. For example, it is possible to start from a large unexplored imagery dataset, to have an analyst select an image of interest, and to return novel images from the dataset that are matched to the mission.

Part of a DARPA ARGUS-IS image of Quantico Marine Corps Base in Virginia (2nd row), with two targeted areas illustrating the six inch resolution of the camera (1st row). *(image courtesy of DARPA)*

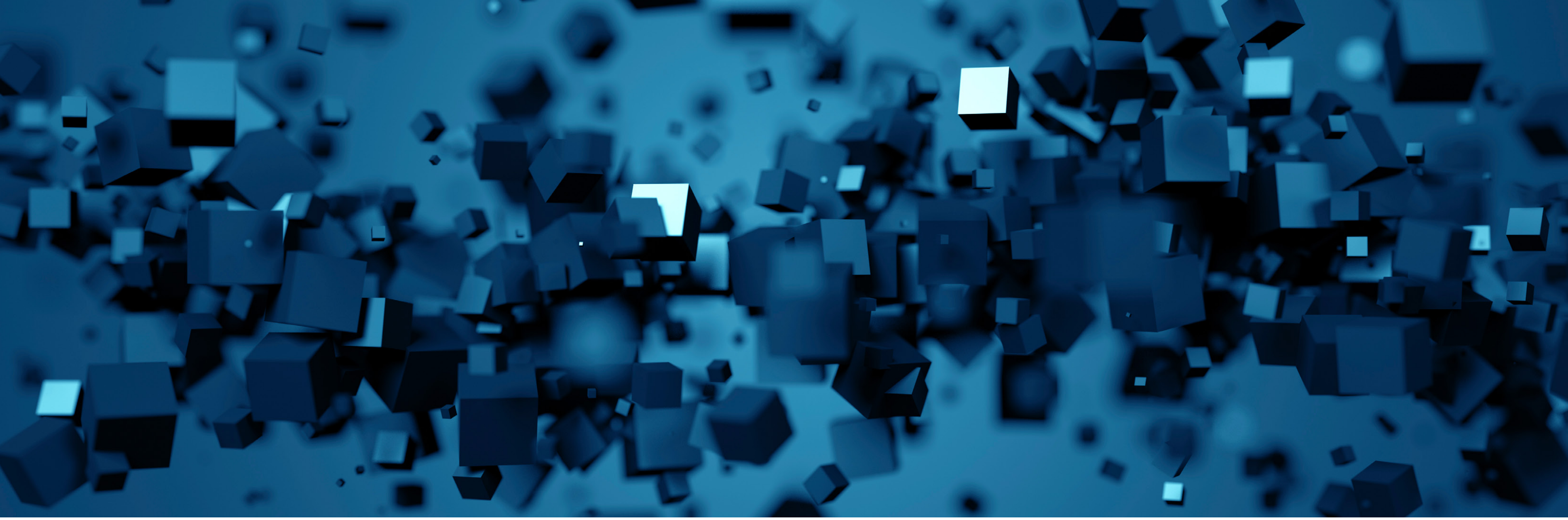


Neural network from brain slice acquired by array tomography, one of the novel image acquisition techniques producing tera-bytes of data a day.

Models

Signal models are fundamental to the question of what information is necessary to obtain and of how that information is applied. Models connect data to decision, providing a common thread that links data acquisition, staging and storage, data analysis and visualization, and magnifies the impact of advances in individual blocks. Models also serve to integrate all the disciplines that play a role in *data science*—from the physics of data acquisition to mathematical methods of data analysis, and to domain-specific expertise.

Advances in compressed sensing have reduced data acquisition volumes by adopting more advanced signal models. They have also brought fundamental change to other blocks in the signal processing chain. The transition from *subspace* models to union of subspace models has resulted in replacement of classical linear methods of data analysis by those based on *convex relaxation*. These developments have breathed new life into *convex optimization*, leading to new algorithms and new geometric methods of performance analysis.



Engineering practice in many important areas has been fundamentally changed by applying the power of convex relaxation to the new *data models*. Examples include fast MRI acquisition, phase retrieval, image denoising, video and motion segmentation, as well as recommendation systems such as the NETFLIX movie recommendation system.

Algorithms

It is necessary that a model is expressive, but it is essential that a model affords efficient computation.

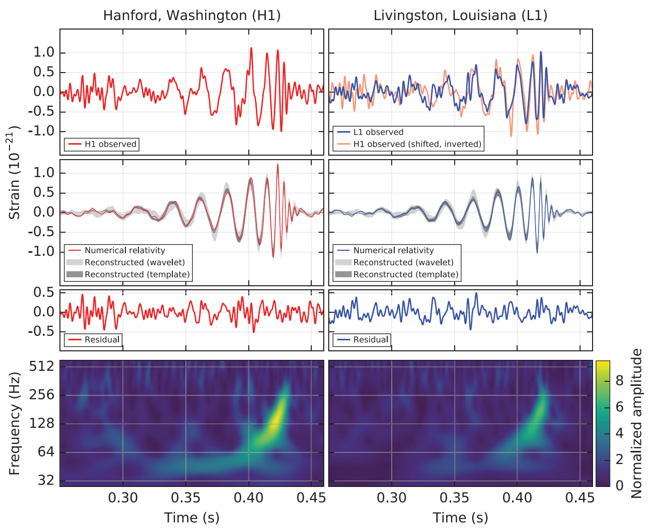
Algorithms provide the computational thread that links *data* acquisition, staging and storage, analysis and visualization. Once this thread is in place, it is possible to start from a data source and some specific knowledge of the application and to provide insight in hours, days or weeks rather than months, years or not at all.

The *data science* community has developed a very successful approach to attacking hard problems: First, find a tractable convex surrogate, second minimize the surrogate, and then prove that for well-structured instances the solution is accurate. It has also developed a suite of extraordinarily effective algorithms; including Interior Point methods, proximal gradient and its variants, augmented Lagrangian and ADMM, re-discovery and extensions of Frank-Wolfe, and efficient constrained matrix factorization.

Architecture

The data to decision pipeline combines information that is present in signals at different scales.

Extraordinary empirical advances in image classification have resulted from the transition of handcrafted features and bag-of-words classifiers to *deep neural networks* that learn hierarchies of features. These advances are motivating the development of new mathematical theory.



The recent direct detection of gravitational waves was enhanced with use of a wavelet signal processing framework developed by Ingrid Daubechies and her colleagues. *[Abbott, et al. 2016 Phys Review Letters, DOI: 10.1103/PhysRevLett.116.061102]*

Conceptual Challenges

The data to decision challenge is illuminating new research directions in information theory, statistics, mathematics, machine learning, and control. For example, information theory was developed for the asymptotic limit of large block lengths. New theory is needed for the finite block length regime. We lack solid understanding of what is possible in the limit of runtime approaching zero, power consumption approaching zero, and data complexity becoming unbounded. We describe some of the critical challenges discussed at the workshop below, following the same template as before.

Signal Acquisition

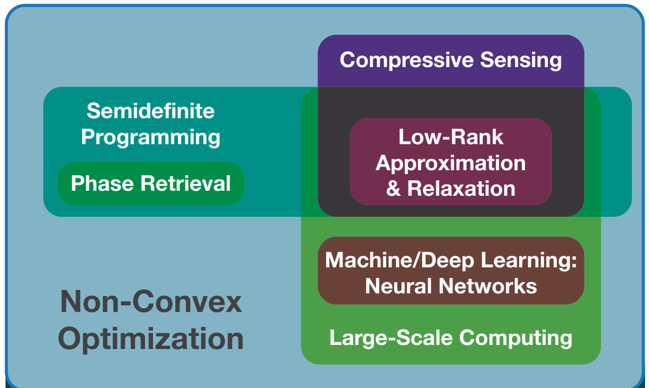
The rate at which digital data is becoming accessible is increasing rapidly. *Data* sets are becoming more massive, more complex, and more difficult to annotate. New data types are emerging related to personal health, social networks, and human behavior, that are challenging our conceptions of privacy and security.

New statistical methods are required to answer the question of when to collect more data and how to weight the data that has been collected. Of particular interest is *adaptive sensing* when multiple modalities are present, where potential value is compromised by the fragility of modeling assumptions, and by errors and biases in data collection. Also of particular interest are small sample regimes with rare events, missing data, and corrupted data.

The challenges associated with *distributed* and *multimodal sensing* need to be resolved. New architectural principles are needed to answer the question of which device will be responsible for sensing, how sensing will be carried out, and how individual results will be combined. New methods of analysis are required that assess value / sample, / joule, / Hz, / flop, / nm, and / bit. New theory is needed that defines and bounds privacy while still enabling distributed processing.

In areas such as face recognition, data will always be highly variable and methods need to account for occlusions and variations in pose and lighting. A first step toward addressing data quality issues is to develop solutions that are specific to a type of deformation associated with a class of applications. Despite extensive research, *analog to digital conversion* remains a roadblock in applications that require high sampling rates and high bandwidth. This leads to high power and high cost that blocks development of new technologies such as the cognitive radio. Developing signal specific methods of analog to digital conversion may provide a way forward.

Non-Convex Problems in Data Science



Dictionary-based methods in CS involve non-convex bi-linear optimization. These kinds of problems are closely related to semidefinite programs (SDP), and in particular to the closely related problem of phase retrieval. SDPs, like dictionary learning problems, are now commonly being solved using non-convex bi-linear approximations rather than full-scale convex semidefinite relaxations. Such non-convex formulations allow problems to be solved at a much larger scale than before, although with weaker guarantees of optimality. The idea of non-convex optimization has been present for a long time in the machine learning community, where it's known that we are able to get good solutions to non-convex problems like deep nets using gradient methods and other heuristics. *(courtesy of Prof. Thomas Goldstein.)*

Models

Convex optimization has become the workhorse of data analysis, but when models are made more expressive by adding in constraints like non-negativity or structure in *matrix factorization*, these models become non-convex, and data analysis becomes less tractable. Understanding the trade off between expressiveness and computational complexity would be a first step in expanding the concept of model-validation to encompass all elements of the pipeline from data to decision.

Algorithms

As models become more expressive, the data landscape becomes non-convex, and new methods are needed to understand and to navigate *local minima*. Many important problems fall outside the scope of known convex methods. A particularly important example is deep learning, where measures of performance that depend on sample complexity and the number of levels in the convolutional network have yet to be developed.

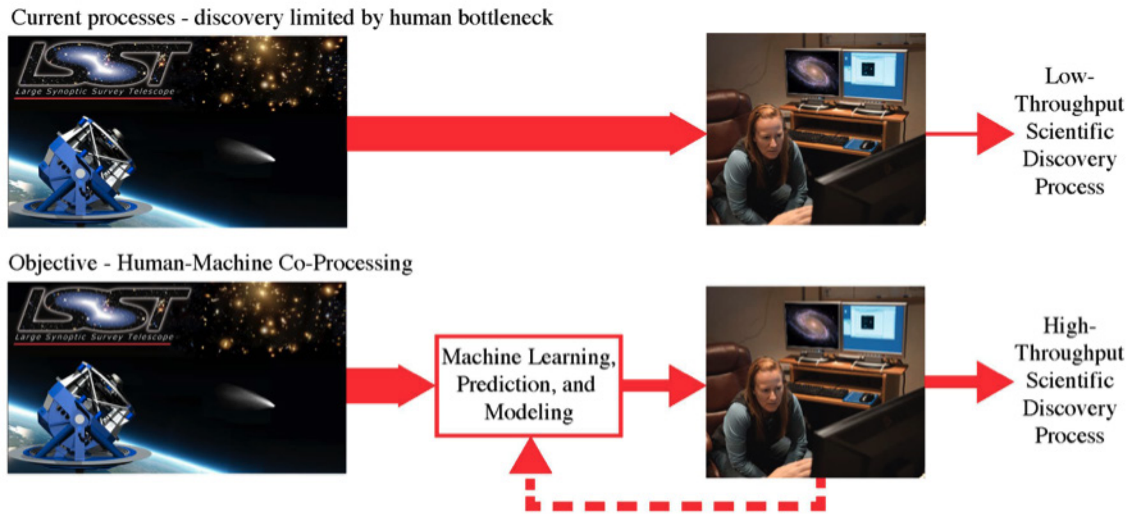
The field of *non-convex optimization* is not yet well understood. The most widely used algorithm is (stochastic) gradient descent; even here we lack methods to quantify the suboptimality of local minima, principles for selecting the starting point, and estimates of what might be gained by exploring alternative starting points. **This interplay between modeling and algorithms demonstrates the importance of managing the *data science* pipeline as a whole.**

The combined talents of humans and computers, when working together in partnership and symbiosis, will be significantly more creative than any computer or human working alone. Integrating human experts into a variety of *data science* pipelines will expand our understanding of how human speed and capacity limits the potential of human-machine co-processing.

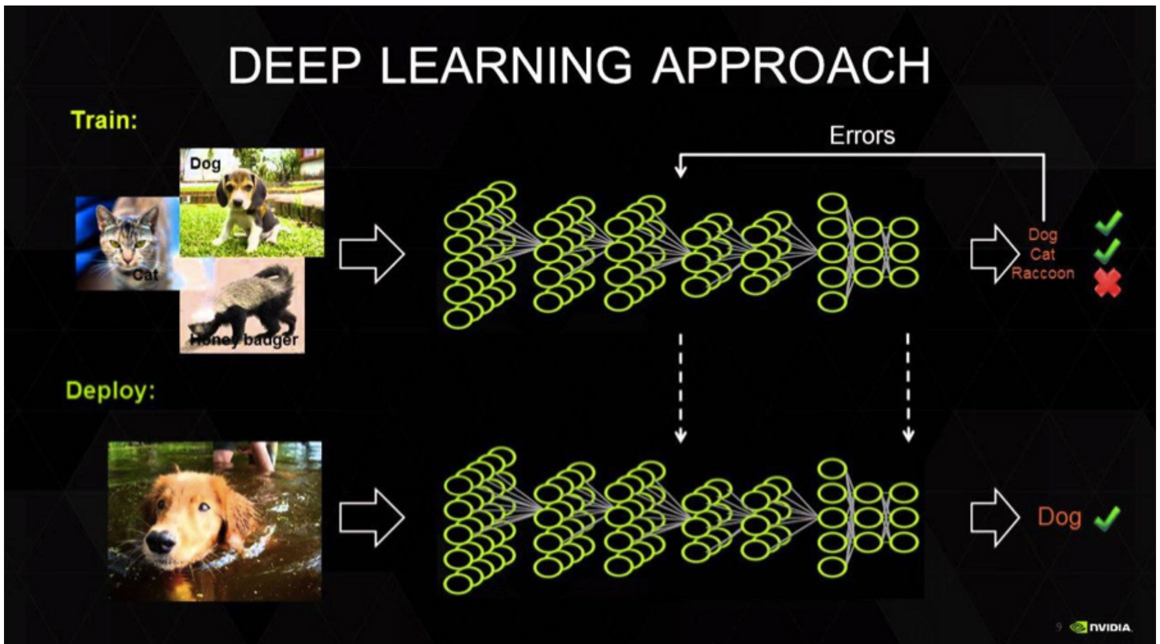
Architecture

Distributed architectures for *data* processing integrate the results of local computations. In applications of deep learning to image classification, local features are combined to produce higher level features. Methods have been developed for reasoning about information flows in communication networks, but we have not yet developed methods of reasoning about equivocation in data processing architecture. A first step might be to develop separation principles such as *spatio-temporal factorization*.

Of particular interest is the challenge of understanding deep learning. As deep convolutional networks achieve more empirical success, it becomes more important to be able to explain why they are effective. It is natural to ask whether more than a hundred layers are necessary, how much training data is required, and whether we should be concerned about false positives. The use of deep learning in mission critical applications will be limited until these questions are adequately addressed.



Integration of human and machine will improve scientific discovery. *(courtesy of Prof. Rebecca Willett)*



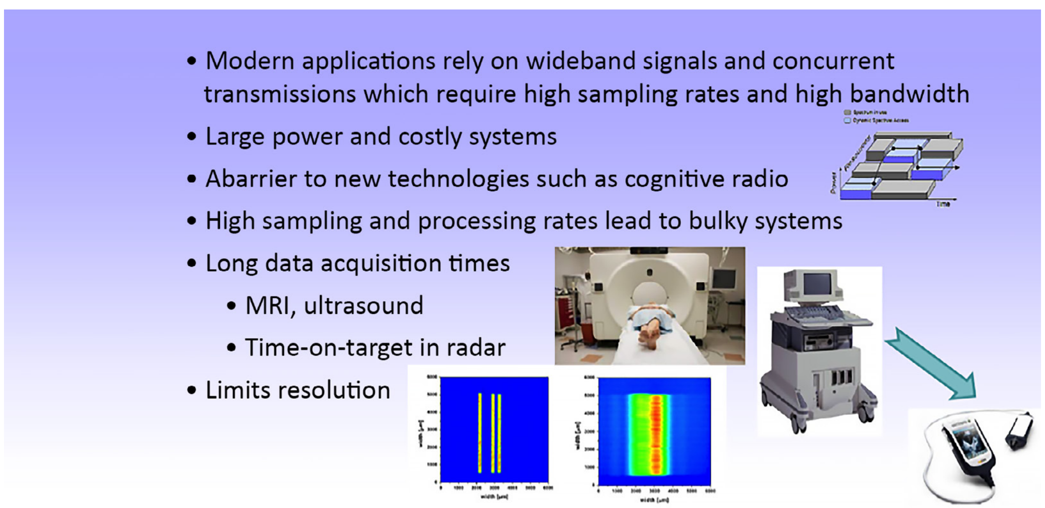
Example of a deep learning architecture for image analysis. *(courtesy of NVIDIA)*

Trajectories for Compressed Sensing Research

We now present specific areas and initiatives that workshop participants considered the trajectory of research necessary to meet *data science* research challenges of future decades. A time horizon (10 or 15 years) is provided to offer a general sense of how close the field is to achieve a specific research goal. The areas are also classified as either Science (S), or Technology (T), or both (S&T), further stressing the main bottlenecks and where the main progress is needed.

10 Year Horizon:

Sensing+X or Task Oriented Sensing (S&T): *Compressed sensing* has focused on reconstruction of generic sparse signals, but new directions with great promise involve the detection of very specific signals and *sensing* of signal attributes. The covariance matrix or power spectrum is one example. Potential applications include development of custom MRI focused on detection of specific pathologies such as brain tumors and cameras just to detect faces. Significant gains in sensing performance can be expected from adoption of more parsimonious goals.



(courtesy of Prof. Yonina Eldar)

Model Driven Signal Acquisition (S&T): Analog-to-digital converters (ADC) remain a major bottleneck. The workshop participants see an opportunity to take advantage of *data models* to meet the challenge of designing systems that are small, cheap, low power, and compact. Furthermore these systems should enable fast acquisition, improved resolution, and recovery of targeted information at significantly reduced bandwidth. These objectives will require co-design of: analog and digital components, the data model, and downstream *optimization*. From a theoretical perspective, the study of these signal processing architectures is an opportunity to unify sampling and rate distortion theory.

What to Model (S): Big data dominates discussion of data science but modeling small data is equally challenging and important for learning parameters. The consequences of model mismatch for performance sensitive applications are not yet fully understood. An emerging challenge is that of modeling the human in the loop, where issues like latency, fatigue, and bias need to be considered.

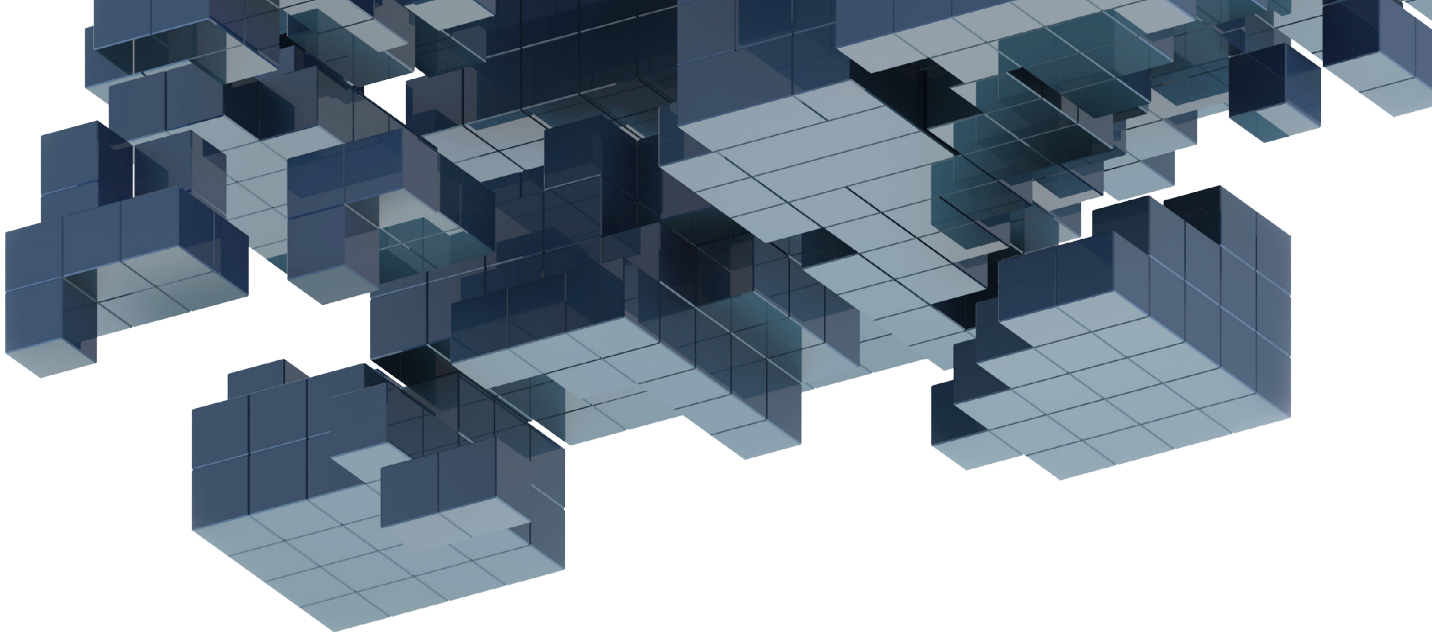
More attention has traditionally been paid to modeling the signal of interest than modeling the background against which detection takes place. However it is much more challenging to detect a person of interest in a subway station than to detect the same person in a photography studio. This suggests an emphasis on modeling both background clutter and signal of interest. Geometric and statistical methods, as well as learning, play a role here.

New developments in sequential experimental design (sequential adaptive) are needed for applications in healthcare and personalized medicine. This is a clear example where sensing can be expensive and task-dependent adaptive and compressive *sensing* is a must.

New Methods of Regularization (S): This includes development of algorithms for capturing salient properties of large data sets in sketches, and development of data-dependent regularizers. The objective here is a *compressive sensing* framework in which computational algorithms are able to extract and take advantage of specific data structure.

The Interface between Modeling and Computation (S): Convexity provides an interface that makes it possible to choose from a class of models and choose from a class of *optimization* algorithms. It has encouraged experimentation on both sides of the interface. New signal representations, new models, and new interfaces (separation principles) are needed to realize the potential of non-linear optimization. The objective here is to develop a *data to decision* pipeline that is able to take full advantage of new methods in optimization.

Again it is important to learn from examples in developing new methods of *regularization*. Some examples should involve humans in the loop.



15 Year Horizon:

Robust Data Fusion (S): A theoretical framework is needed for robust data fusion to support data acquisition from multiple sources and modalities and to quantify uncertainty in such data. Models that capture the geometry of data are needed to describe complex sensing environments that are heterogeneous, dynamic, and multiscale. The objective here is a compressive sensing framework that encompasses adaptive sensors that manage different local tasks. This framework should take advantage of opportunities presented by processing delay and by the multimodal and heterogeneous data.

Development of Non-Convex Optimization (S): We need to understand what types of structure we can recover with guarantees, and how these structures might be combined. Factorization is a path to divide and conquer solutions and matrix factorization might be a good benchmark problem for non-convex methods.

How to Predict Performance (S): Some algorithms and models empirically appear optimal for large datasets (e.g., *deep neural networks*) and some for small datasets (e.g., random forests and *sparse modeling*). The objective here is to develop a compressive *sensing* and data analysis framework that predicts performance for all sizes of data.

It is important to listen to the success of ad-hoc techniques (deep learning being one of them). The challenge is to understand empirical success. By working on the fundamentals, we should be able to not only understand why ad-hoc techniques work, but also how to improve them. To this end, it might be helpful to derive negative examples, including some where there is a human in the loop and some where limited training data is available.

How to Measure Performance of Algorithms (S): The objective here is a framework for performance analysis that encompasses new methods of optimization and new computer architectures for implementation. It is important to match performance measurement to task, e.g., go beyond often used average performance metrics that are not useful for critical high-risk operations.

Development of dual certificates for convex relaxations is a promising path to upper bounds on performance.

Average/mean performance over a large testing sample may not always be the most appropriate metric. In medicine and autonomous vehicles, the number of false positives is certainly more critical. Theory is needed for more challenging performance metrics.

There is some evidence to suggest that certain *non-convex optimization* problems become tractable when the input data are large and random. We need to understand the extent to which we can simply apply efficient heuristics to data without worrying about convergence.

Algorithms also need to take into account the computer architecture where they will be implemented. More interaction is needed between modeling and algorithms and hardware teams.

Development of Architectural Principles (S): We need to develop methods of integrated sensing and processing that parallel communication networks. We need to understand what the results of local processing should be, and how they should be combined to accomplish a global inference objective. We need to understand the advantages of active learning in finding an optimal classifier with less human assistance. Finally we need to incorporate processing constraints such as privacy and time bounded computation, energy, and communication.

Investments in Infrastructure

The theoretical and algorithmic developments mentioned before can't happen in a vacuum. There is a clear need to invest in infrastructure that will support such developments. The workshop participants discussed the infrastructure necessary to meet the challenge.

Datasets and Competitions

Datasets such as FERET, MNIST, COIL, YaleFaceB, Caltech, PASCAL, and ImageNet have proved fundamental both to evaluating progress and stimulating development of new theory and algorithms. These datasets have mostly appeared in image processing and computer vision, though exist in other areas such as audio processing. Designing open datasets in other areas, and maintaining them dynamically, is challenging, but a worthwhile infrastructure investment. It is important to note that those datasets do not necessarily need to be shared, in case for example of privacy or confidentiality, the community can access them via virtual machines for algorithm testing. The community, industry, or the government can own the datasets.

It is interesting to note that the above datasets not only advanced the research in a given application, but also opened the door to new or significantly increased activities in others—image processing and analysis are examples. Research from the sixties to the nineties focused on local tasks, such as Image Denoising, image inpainting, image compression, edge detection, and deconvolution. Since the nineties, the focus has been high-level tasks. The high-level tasks research began with pattern recognition at different scales, and then moved to general object classification, following by captioning of image and video sources. While *data* and standard datasets are not the only reason for this evolution, they have certainly become core contributors.

Closely related to this is the topic of organizing open competitions for major challenges. Among the most famous are the NETFLIX movie recommendations competition and multiple DARPA robotics challenges. Some of these challenges were instrumental in development of self-driving vehicles. Other areas such as protein folding have been running competitions very successfully. The same can be said for the ImageNet competition in computer vision. There may be disagreement on the ideal format for these competitions, but there is complete agreement on their value in advancing knowledge. Promoting design competitions that challenge the community to find the best way to sense certain data to achieve a specific pre-defined task is recommended.

Competitions can be long-term (10-15 years) grand challenges with intermediate targets. Such long-term challenges will also encourage collaborations to form along the way. Incorporating a specific task, such as automatic screening of a particular disorder, will encourage development of end-to-end solutions that combine *sensing*, machine learning, and domain-specific performance metrics.

Competitions to develop new algorithms that are fundamental to all *data science* applications would have a broad impact. For example, it will be outstanding to have by 2030 an algorithm that can multiply two arbitrary n×n matrices in order $n^2 \log(n)$ operations. Admittedly, this would not only have a major impact on *compressed sensing*, but on the entire field of numerical linear algebra (since it would give rise to SVDs, QRs, etc. with the same complexity) and for Scientific Computing at large. A competition to achieve this capability would drive research in this area and potentially encourage collaboration.

Areas like radar signal processing lack benchmark datasets against which new algorithms can be measured. Development of common infrastructure might encourage a community of theorists and experimental researchers that listen and learn from each other.

Open Source Data Science

Long ago, the world had to program in very low-level hardware oriented language. Consequently, programming was restricted to specialists. Programming has evolved, and more and more advanced languages have been developed, from Pascal to C++ to MatLab, opening the door to non-experts to program some of the most extremely sophisticated hardware and to develop complex algorithms. One could consider the same happening with algorithms in the area of data science. Numerous efforts in academia and industry are geared towards making data science transparent to the user, or at least building the basic infrastructure such that data science becomes often like building Lego constructions (APIs for data science). Investing in such a direction will make significant progress in the applications of data science and the integration of sensing and processing in particular. Projects like NEXT (University of Wisconsin)—an *open source* cloud-based system for integrated sensing and processing and active machine learning—removes the burden of GUI, experiments, database, and the computational backend from the researcher, while reducing barriers to entry.

International Centers of Research

Companies like Facebook and OpenAI have invested orders of magnitude more in AI than has the US government.

Google, Facebook and Microsoft constitute International Centers of Excellence. Industry is motivated by self-interest rather than the national interest, but finding ways to collaborate would accelerate the translation of theory to practice and would enhance training of graduate students. Today academia is losing talent to industry, and though this can be viewed positively, the flow of talent cannot be sustained indefinitely.

Academic Centers of Excellence would provide a counterweight. They would bring a focus on emerging applications, such as mobile applications for community health. They would also provide opportunities for interdisciplinary teams to form and develop end-to-end solutions. One could also imagine opportunities to partner with national laboratories on applications related to energy and the environment.

The recently created Turing Institute, created by the UK government with significant industrial collaboration, is one center that moves in this direction and can serve both as a learning example and an international partner.

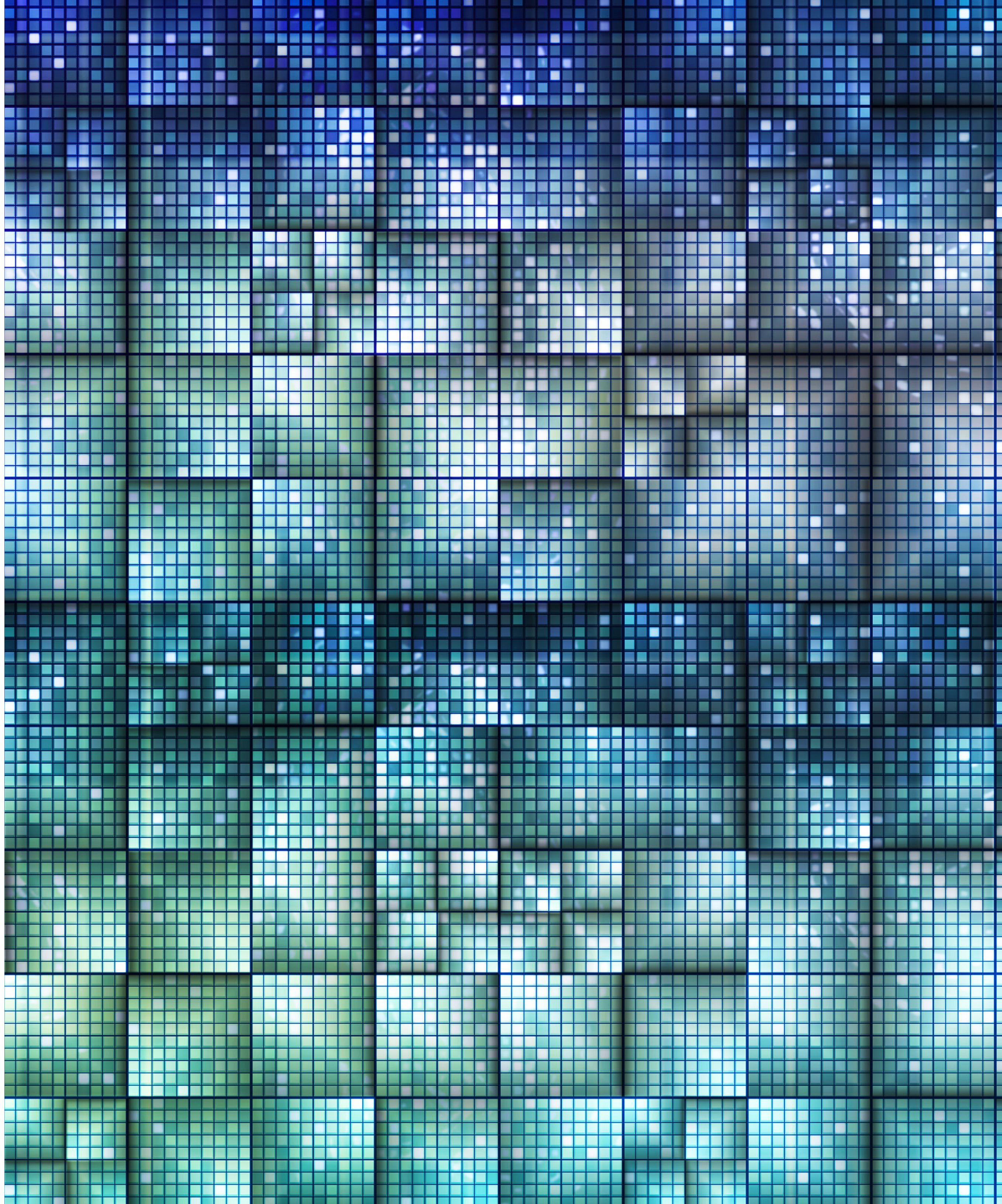
Photo of The British Library where the Alan Turing Institute is headquartered.

Conclusions

Looking ahead to 2030, we are certain to be drowning in ***data*****,** even if we are only able to imagine a fraction of the data types. In spite of initial success stories, understanding such data, developing models and algorithms to analyze data, and processing the data in a resources-efficient fashion are all challenges that go way beyond current available tools.

This report describes a trajectory of research that focuses on new data models, optimization techniques, performance metrics and architecture designs to meet these challenges in next 10 and 15 years. Participants expect these developments to have significant applicability for today's emerging applications like personal health monitoring, automated transportation, social networks and group behavior. Each new application illuminates the research frontier in data science, and not always in the same way.

Participants also outlined the infrastructure necessary to support the research development, like open and benchmark datasets and new competitions that can drive innovation. They conclude that long-term investment is needed to ensure that the research meets its potential and avoids being outflanked by new developments. The return on investment for individual elements should be measured in terms of the effectiveness of the end-to-end data to decision pipeline.



Glossary

Adaptive sensing – Sensing actions are guided by information gleaned from previous measurements. 7, 9

Analog to digital converters – A device that converts a continuous physical quantity to a digital number that represents the amplitude of the quantity. 9

Compressed sensing (alt. compressive sensing) – A method of capturing attributes of a signal with very few measurements. 6, 7, 11, 13

Convex function – A continuous function whose value at the midpoint of every interval in its domain does not exceed the arithmetic mean of its values at the ends of the interval.

Convex optimization – The problem of minimizing a convex function over a convex set. 7, 9

Data – A set of values of qualitative or quantitative variables. 4, 5, 6, 7, 8, 9, 10, 11, 13, 15

Data compression – Encoding information with fewer bits than the original representation.

Data models – Organize data elements and standardize how the data elements relate to one another. 11

Data science – An interdisciplinary field dealing with processes and systems that extract information from data in various forms. 6, 8, 9

Deep neural networks – A set of algorithms that attempt to model high-level abstractions in data using multiple processing layers that employ non-linear transformations. 8, 12

Distributed sensing – sensing done with multiple sources

Epigenetic – Relating to or arising from non-genetic influences on gene expression. 3

Global minima – The smallest overall value of a function over its entire range.

Local minima – The minimum of a function within some neighborhood. 9

Multimodal sensing – sensing that includes more than one modality, for example RGB and infrared data or video for audio data. 9

Non-convex function – A function that is not convex.

Non-convex optimization – The problem of minimizing a non-convex function over a given set. 9, 12

Non-negative matrix factorization – A group of algorithms where a matrix V is written as the product of two matrices W and H, with the property that all three matrices have no negative elements. 9

Open source – Software for which the source code is exposed and freely available. 5, 13

Optimization – The selection of a best element from some set of available alternatives. 11

Regularization – The process of introducing additional information in order to solve and ill-posed problem or to prevent over fitting. 11

Sensing – Detecting events or changes in an environment and then providing a corresponding output. 5, 11, 12, 13

Smart sensing – sensing guided by information, task and decisions. 4

Sparse model – Organization in a lower dimensional space of values observed in high dimensional space. 12

Spatio-temporal factorization – Factorizing a function in terms of spatial and temporal factors or basis. 10

Subspace – A subset of a vector space that is closed under addition and scalar multiplication. 7

Supervised learning – The machine learning task of inferring a function from labeled training data.

Union of subspaces model – Signal coefficients that lie in certain subspaces are active or inactive together. The potential subspaces are known in advance, but the particular set of subspaces that are active in the signal support must be learned from measurements.

Unsupervised learning – The machine learning task of inferring a function to describe hidden structure from unlabeled data.

Appendix I
Compressive Sensing Researchers

Helmut Bolcskei
ETH Zurich, boelcskei@nari.ee.ethz.ch
Communication Technology Laboratory
PhD (1997), Electrical Engineering, Vienna University of Technology

Dr. Bolcskei's research interests are in information theory, mathematical signal processing, machine learning, and statistics.

He has received the 2001 IEEE Signal Processing Society Young Author Best Paper Award, the 2006 IEEE Communications Society Leonard G. Abraham Best Paper Award, the 2010 Vodafone Innovations Award, the ETH "Golden Owl" Teaching Award, is a Fellow of the IEEE, a 2011 EURASIP Fellow, a 2013-2014 Distinguished Lecturer of the IEEE Information Theory Society, and was an Erwin Schrödinger Fellow (1999-2001) of the Austrian National Science Foundation (FWF). Dr. Bolcskei has been a plenary speaker at several IEEE conferences and served as an associate editor of the IEEE Transactions on Information Theory, the IEEE Transactions on Signal Processing, the IEEE Transactions on Wireless Communications, and the EURASIP Journal on Applied Signal Processing. He was editor-in-chief of the IEEE Transactions on Information Theory during the period 2010-2013, and serves on the editorial boards of "Foundations and Trends in Networking", "Foundations and Trends in Communications and Information Theory", and the IEEE Signal Processing Magazine.

Alex Bronstein – *http://www.cs.technion.ac.il/~bron/*
Technion, Israel Institute of Technology, bron@cs.technion.ac.il
Department of Computer Science
PhD (2007), Computer Science, Technion.

Dr. Bronstein's main research interests are theoretical and computational methods in metric geometry and their application to problems in computer vision, pattern recognition, shape analysis, computer graphics, imaging and image processing, and machine learning. He has authored over 120 publications in leading journals and conferences, over two dozen patents and patent applications, and the book Numerical geometry of non-rigid shapes.

Bronstein's research has been recognized by numerous awards, including the Kasher prize (2002), Thomas Schwartz award (2002), Hershel Rich Technion Innovation award (2003), Gensler counter-terrorism prize (2003), the Copper Mountain Conference on Multigrid Methods Best Paper award (2005), the Adams Fellowship (2006), the Krill Prize by Wolf Foundation (2012), and the European Research Council (ERC) Startup Grant (2013). Highlights of his research have been featured on CNN and SIAM News.

Robert Calderbank – *http://ece.duke.edu/faculty/robert-calderbank*
Duke University, robert.calderbank@duke.edu
Department of Computer Science
PhD (1980), Mathematics, California Institute of Technology

Dr. Calderbank is the Charles S. Sydnor Professor of Computer Science at Duke University. Previous to joining Duke in 2010, he was a Professor of Electrical Engineering and Mathematics and Princeton University. In addition to his teaching background, Calderbank has been the Vice President for Research at AT&T and Bell Labs. At Bell Labs, Calderbank developed voiceband modem technology that was widely licensed and incorporated in over a billion devices. Together with Peter Shor and colleagues at AT&T Labs, Dr. Calderbank developed the group theoretic framework for quantum error correction. This framework changed the way physicists view quantum entanglement, and provided the foundation for fault tolerant quantum computation.

Dr. Calderbank has also developed technology that improves the speed and reliability of wireless communication by correlating signals across several transmit antennas. Invented in 1996, this space-time coding technology has been incorporated in a broad range of 3G, 4G, and 5G wireless standards.

Dr. Calderbank is an IEEE Fellow and an AT&T Fellow, and he was elected to the National Academy of Engineering in 2005. He received the 2013 IEEE Hamming Medal for contributions to coding theory and communications and the 2015 Shannon Award.

Alexey Castrodad
National Geospatial-Intelligence Agency

Yonina Eldar – *http://webee.technion.ac.il/people/YoninaEldar/index.php*
Technion, Israel Institute of Technology, yonina@ee.technion.ac.il
Department of Electrical Engineering
PhD (2002), Electrical Engineering and Computer Science, MIT

Dr. Eldar's research interests include technology development in the areas of signal processing, medical imaging, communications, sampling and ADC design, signal processing for optics and biology. Every year for the last 15 years of her career, she has received at least one accolade or award for her research. Most recently in 2016, she was awarded the IEEE Kiyo Tomiyasu Award "for development of the theory and implementation of sub-Nyquist sampling with applications to radar, communications, and ultrasound." In addition to her research, Dr. Eldar has been recognized as a distinguished lecturer, and in 2011, was selected as one of the 50 most influential women in Israel.

Anna Gilbert
University of Michigan, annacg@umich.edu
Department of Mathematics
PhD (1997), Mathematics, Princeton University

Dr. Gilbert's research interests include analysis, probability, networking, and algorithms. She is particularly interested in randomized algorithms with applications to harmonic analysis, signal and image processing, networking, and massive datasets. In addition to her academic background, Dr. Gilbert has worked as a technical staff member at AT&T Labs-Research in Florham Park, NJ (1998-2004). She has received several awards, including a Sloan Research Fellowship (2006), an NSF CAREER award (2006), the National Academy of Sciences Award for Initiatives in Research (2008), the Association of Computing Machinery (ACM) Douglas Engelbart Best Paper award (2008), the EURASIP Signal Processing Best Paper award (2010), and the SIAM Ralph E. Kleinman Prize (2013).

Andrea Goldsmith
Stanford University, andrea@ee.stanford.edu
Department of Electrical Engineering
PhD (1994), Electrical Engineering, University of California Berkeley

Dr. Goldmisth's research interests include wireless information and communication theory, cognitive radios, sensor networks, "green" wireless system design, control systems closed over wireless networks, smart grid sensing and control, and applications of communications and signal processing to biology and neuroscience.

Andrea Goldsmith is the Stephen Harris professor in the School of Engineering and a professor of Electrical Engineering at Stanford University. She co-founded and serves as Chief Scientist of Accelera, Inc., which develops software-defined wireless network technology, and previously co-founded and served as CTO of Quantenna Communications Inc., which develops high-performance WiFi chipsets. She has previously held industry positions at Maxim Technologies, Memorylink Corporation, and AT&T Bell Laboratories. Dr. Goldsmith is a Fellow of the IEEE and of Stanford, and she has received several awards for her work, including the IEEE Communications Society and Information Theory Society joint paper award, the IEEE Communications Society Best Tutorial Paper Award, the National Academy of Engineering Gilbreth Lecture Award, the IEEE ComSoc Communications Theory Technical Achievement Award, the IEEE ComSoc Wireless Communications Technical Achievement Award, the Alfred P. Sloan Fellowship, and the Silicon Valley/San Jose Business Journal's Women of Influence Award.

Tom Goldstein
University of Maryland, tomg@cs.umd.edu
Department of Computer Science
PhD (2010). Applied Mathematics, University of California Los Angeles

Dr. Goldstein's research focuses on the intersection of optimization, machine learning, distributed computing, and image processing. He is interested in fast, low-complexity solutions to real-world model-fitting problems from data analytics and imaging. He seeks to develop massively distributed methods for analyzing big data, as well as simple efficient schemes for small embedded platforms. Dr. Goldstein is a recent Richard C. DiPrima Prize winner, which is awarded to one person every two years by the Society for Industrial and Applied Mathematics (SIAM) for outstanding research in applied mathematics.

Ronnie Hadani
University of Texas, hadani@math.utexas.edu
Department of Mathematics
PhD (2006), Mathematics, Tel-Aviv University

Dr. Hadani's research interests include Representation Theory, Theory of Algebraic D-modules, and applications to harmonic analysis, signal processing, three dimensional cryo-electron microscopy, and mathematical physics. He has recently been granted several patents, including "finite crystal oscillator" and "communications method employing orthonormal time-frequency shifting and spectral shaping." He is also a co-founder of Cohere technologies Inc., a company that focuses on the development of advanced communication technologies.

Babak Hassibi
California Institute of Technology, hassibi@systems.caltech.edu
Department of Electrical Engineering
PhD (1996), Electrical Engineering, Stanford University

Dr. Hassibi's research is in communications, signal processing, and control. He is currently interested in wireless networks and in genomic signal processing. In the wireless network area, Hassibi studies modeling issues, information-theoretic questions, scheduling, protocols, and various performance criteria, etc. In the genomic signal processing area he researches real-time DNA microarrays, a novel technology that his lab has developed.

Babak Hassibi is a Gordon M. Binder/Amgen Professor. He has received many awards, including the "Al-Marai Award for Innovative Research in Communication (2009)," "Presidential Early Career Award for Scientists & Engineers (2003)," and a National Science Foundation CAREER award (2002). In addition to his awards, he is an Associate Editor for IEEE Transactions on Information Theory, an editorial board member for the Foundations and Trends in Communications and Information Theory, and an ONR Communications and Technology Review panelist.

Michael Lustig – *http://www.eecs.berkeley.edu/~mlustig/mlustig-cv.pdf*
University of California Berkeley, *mlustig@eecs*
Department of Electrical Engineering & Computer Sciences
PhD (2008), Electrical Engineering, Stanford University

Michael Lustig's research focuses on medical imaging, particularly Magnetic Resonance Imaging (MRI), and very specifically, the application of compressed sensing to rapid and high-resolution MRI, MRI pulse sequence design, medical image reconstruction, and inverse problems in medical imaging and sparse signal representation. Dr. Lustig was the first to develop and demonstrate the application of compressed sensing to rapid MRI. Additionally, he developed parallel imaging reconstruction techniques, image artifacts reduction methods, motion estimation and correction techniques, novel RF-excitation pulses, rapid MRI pulse sequence design, and improved functional MRI techniques. Among other accolades, Dr. Lustig is the GE Healthcare Thought Leader recipient for "ground-breaking work in compressed-sensing MRI."

Gonzalo Mateos – *http://www.ece.rochester.edu/~gmateosb/cv.html*
University of Rochester, *gmateosb@ece.rochester.edu*
Department of Electrical Engineering and Computer Science
PhD (2012), Electrical Engineering, University of Minnesota

Dr. Mateos' research is on algorithms and analysis; specifically applications of statistical signal processing tools to dynamic network health monitoring, social, power grid, and Big Data analytics. His current research includes robust, distributed, and sparsity-aware learning from high dimensional social data and spectrum sensing for wireless cognitive radio networks. Having recently attained his PhD, Dr. Mateos' thesis was awarded the "Best Dissertation Award Honorable Mention." The title of his submission was "Sparsity Control for Robustness and Social Data Analysis." Mateos currently teaches Data Science and Stochastic Systems at the University of Rochester.

Andrea Montanari
Stanford University, *montanari@stanford.edu*
Department of Electrical Engineering & Department of Statistics
PhD (2001), Theoretical Physics, Scuola Normale Superiore (Italy)

Dr. Montanari's research interests include machine learning, high-dimensional statistics, graphical models, coding theory, and random combinatorial structures and optimization. He was co-awarded the ACM SIGMETRICS best paper award in 2008. He received the CNRS bronze medal for theoretical physics in 2006, the National Science Foundation CAREER award in 2008, the Okawa Foundation Research Grant in 2013, and the Applied Probability Society Best Publication Award in 2015. He is an Information Theory Society distinguished lecturer (2015-2016).

Robert Nowak
University of Wisconsin, *nowak@ece.wisc.edu*
Department of Electrical and Computer Engineering
PhD (1995), Computer Science, University of Wisconsin

Dr. Nowak's research interests include signal and information processing, machine learning, optimization, and statistics. Recently, Nowak has developed two highly successful apps in BeerMapper and NEXT that apply findings from his research.

He has earned numerous awards and accolades including the ECML-PKDD award for best paper on Knowledge Discovery (2012), the Grand Award recipient of the Talbert Abrams Paper Award (2012) and the ERDAS Award for Best Scientific Paper in Remote Sensing (2012).

Henry Pfister
Duke University, *henry.pfister@duke.edu*
Department of Electrical and Computer Engineering
PhD (2003), Electrical Engineering, UCSD

Dr. Pfister's current research interests include information theory, channel coding, and iterative information processing with applications in wireless communications, data storage, and signal processing.

He has received the NSF Career Award in 2008, the Texas A&M ECE Department Outstanding Professor Award in 2010, and was a coauthor of the 2007 IEEE COMSOC best paper in Signal Processing and Coding for Data Storage. He is currently an associate editor in coding theory for the IEEE Transactions on Information Theory (2013-2016).

Ben Recht – *http://www.eecs.berkeley.edu/~brecht/bio.html*
University of California Berkeley, *brecht@berkeley.edu*
Department of Electrical Engineering and Computer Science
PhD (2006), Applied Mathematics, MIT

Dr. Recht's research interests include trying to find mathematical solutions to data analysis that bridge across scientific fields and applications. He is interested in devising algorithms that handle "noisy" and incomplete data.

Ben is the recipient of a Presidential Early Career Awards for Scientists and Engineers (2014), an Alfred P. Sloan Research Fellowship (2012), the 2012 SIAM/MOS Lagrange Prize in Continuous Optimization, the 2014 Jamon Prize, and the 2015 William O. Baker Award for Initiatives in Research. He is currently on the Editorial Board of the Journal for Machine Learning Research.

Galen Reeves
Duke University, *galen.reeves@duke.edu*
Department of Electrical and Computer Engineering
PhD (2011), Electrical Engineering and Computer Science, Duke University

Dr. Reeve's research interests lie at the intersection of signal processing, statistics, and information theory, with applications in compressed sensing, robust statistics, massive data storage and retrieval, neuroscience, and machine learning. He believes that with the "information glut upon us, many of the most important scientific and technological advances of the next several decades will follow from our ability to collect, understand, and communicate massive amounts of data." An overall theme in his research is to draw upon mathematical tools from a wide variety of disciplines -- such as random matrix theory, convex optimization, statistical decision theory, and statistical physics -- to understand the limits of what is possible (and what is impossible) in problems of high-dimensional statistical inference, and also to figure out how to reach these limits using computationally practical methods. His Ph.D. dissertation used tools from information theory to provide a sharp characterization of the problem of sparsity pattern recovery in compressed sensing.

Tom Richardson – *http://ethw.org/Thomas_J._Richardson*
Qualcomm
Qualcomm R&D

Considered one of the world's foremost experts on iterative decoding, Thomas J. Richardson, in conjunction with Rüdiger Urbanke, helped optimize data transmission rates for wireless and optical communications and digital information storage. To approach "Shannon's limit," which established the maximum rate for communications over a noisy channel, they expanded on low density parity check (LDPC) codes and provided new tools for understanding the complexities of iterative decoding procedures. The result has been reliable data transmission at rates close to channel capacity but with low complexities. Three landmark papers by Drs. Richardson and Urbanke appearing in the February 2001 issue of the IEEE Transactions on Information Theory successfully addressed the obstacles facing the development of capacity-approaching codes. Their work showed that LDPC codes could very closely approach the Shannon limit, showed how to design irregular LDPC codes and provided methods for efficiently encoding LDPC codes and they introduced the density evolution technique, on which most subsequent work on LDPC codes is based.

Justin Romberg
Georgia Institute of Technology, *jrom@ecegatech.edu*
School of Electrical and Computer Engineering
PhD (2004), Electrical Engineering, Rice University

Dr. Romberg's research interests include statistical learning and signal processing.

In 2008 he received an ONR Young Investigator Award. In 2009 he received a PECASE award and a Packard Fellowship, and in 2010 he was named a Rice University Outstanding Young Engineering Alumnus. In 2006-2007 he was a consultant for the TV show "Numb3rs" and from 2008-2011, he was an Associate Editor for the IEEE Transactions on Information Theory. He is currently on the editorial board for the SIAM Journal on Imaging Science.

Aswin Sankaranarayanan
Carnegie Mellon University, *saswin@andrew.cmu.edu*
Electrical Engineering Department
PhD (2009), Electrical Engineering, University of Maryland

Dr. Aswin Sankaranarayanan's research interests are in computer vision and signal processing. Specifically, his research focuses on developing computational tools and imaging architectures for high-dimensional visual signals—this encompasses ideas across multiple disciplines: compressive sensing, sparse approximations, multi-view geometry, computational imaging, non-linear signal models and reflectance properties of materials.

Sankaranarayanan has received several awards including the Distinguished Dissertation Fellowship at the University of Maryland (2008-2009), Best Paper Award (2010), and the Future Faculty Fellowship Award (2009).

Guillermo Sapiro
Duke University, *guillermo.sapiro@duke.edu*
Electrical and Computer Engineering
PhD (1993), Electrical Engineering, Technion University (Israel)

Dr. Sapiro works on theory and applications in computer vision, computer graphics, medical imaging, image analysis, and machine learning. He has authored and co-authored over 300 papers in these areas.

Sapiro was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Post-Doctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the National Science

Foundation Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010. He received the test of time award at ICCV 2011.

Phil Schniter
Ohio State University, schniter.1@osu.edu
Electrical and Computer Engineering
PhD (2000), Philosophy, Cornell

Dr. Schniter's research interests are in signal processing for communication systems, adaptive filtering, estimation theory, blind equalization, and identification.

He has received the Lumley Research Award from the college of engineering at Ohio State (2005), Best Paper Award IEEE SPAWC conference (2005), the NSF Career Award (2003), and was granted an Intel Foundation Fellowship (1998-1999).

In addition to his research, he has given numerous talks and presentations regarding sparse reconstruction, graphical model approaches to compressive inference, and advanced equalization techniques for wireless links.

Aarti Singh
Carnegie Mellon University, aartisingh@cmu.edu
Department of Machine Learning
PhD (2008), Electrical and Computer Engineering, University of Wisconsin

Dr. Singh's research goal is to "understand the fundamental tradeoffs between [computational efficiency and statistical optimality], and design algorithms that can learn and leverage inherent structure of data in the form of clusters, graphs, subspaces and manifolds."

She is investigating how the tradeoffs between computational efficiency and statistical optimality can be further improved by designing interactive algorithms that employ judicious choice of where, what, and how data is acquired, stored, and processed. Her vision is to introduce a new paradigm of intelligent machine learning algorithms that learn continually via feedback and make high-level decisions in collaboration with humans, thus pushing the envelope of automated scientific and social discoveries.

Dr. Singh's research has been supported by grants from NSF, AFOSR, and NIH's MIDAS Center at University of Pittsburgh, including NSF CAREER (2013) and BIG DATA awards and the AFOSR Young Investigator Award (2014).

Amit Singer
Princeton University, amits@math.princeton.edu
Department of Mathematics

Dr. Singer's current research is focused on developing algorithms for three-dimensional structuring of macromolecules using cryo-electron microscopy.

Mathematical interests: linear and non-linear dimensionality reduction of high dimensional data, signal, and image processing, spectral methods, convex optimization, and semidefinite programming.

Professor Amit Singer has received the 2010 Presidential Early Career Award for Scientists and Engineers (PECASE), the highest honor bestowed by the U.S. government on science and engineering professionals in the early stages of their research careers. He is among 94 researchers at American institutions selected by the Office of Science and Technology Policy within the Executive Office of the President based on the recommendations of 16 federal departments and agencies. Singer also received a 2010 Sloan Research Fellowship for his research.

Thomas Strohmer
University of California, Davis, lastname@math.ucdavis.edu
Department of Mathematics
PhD (1998), Mathematics, Universität Wien (Austria)

Dr. Strohmer's research interests are in applied harmonic analysis, mathematics of information, numerical algorithms, mathematical signal and image processing, and high-dimensional data analysis.

He has recently published work titled "Accurate Detection of Moving Targets Via Random Sensor Arrays and Kerdock Codes," "Remote Sensing Via Minimization" (Foundations of Computational Mathematics), and contributed to the Foreword, along with Dr. Yonina Eldar, in the IEEE Trans. Aerospace and Electronic Systems issue 50.

Joel Tropp
Caltech, jtropp@cms.caltech.edu
Engineering & Applied Science
PhD (2004), Computational and Applied Mathematics

Dr. Tropp's research interests include randomized algorithms for matrix analysis, architectures and algorithms for compressive sampling, matrix nearness problems, and data analysis.

Tropp was recently awarded the SPIE Compressive Sampling Pioneer Award (2015) and Thomson Reuters Highly Cited Researcher in Computer Science (2014). He was recently invited to provide an address at the SIAM Joint Mathematical Meetings (2015). In addition to these awards and distinctions, he has won numerous best paper awards, is a Sloan fellow, and a Presidential Early Career Award for Scientists and Engineers (PECASE) awardee (2008).

Rene Vidal
Johns Hopkins University, *rvidal@jhu.edu*
Department of Biomedical Engineering
PhD (2003), Electrical Engineering and Computer Sciences, University of California Berkeley

Dr. Vidal's research areas include biomedical image analysis, computer vision, machine learning, dynamical systems theory, and robotics. He is particularly interested in the development of mathematical methods for the interpretation of high-dimensional data, such as images, videos, and biomedical data. He has developed methods from algebraic geometry, sparse and low-rank representation theory for clustering and classification of high-dimensional data, and methods from dynamical systems theory for modeling and comparison of time series data.

Dr. Vidal is recipient of numerous awards for his work, including the 2012 J.K. Aggarwal Prize for "outstanding contributions to generalized principal component analysis (GPCA) and subspace clustering in computer vision and pattern recognition," the 2012 Best Paper Award in Medical Robotics and Computer Assisted Interventions (with Benjamin Bejar and Luca Zappella), the 2011 Best Paper Award Finalist at the Conference on Decision and Control (with Roberto Tron and Bijan Afsari), the 2009 ONR Young Investigator Award, the 2009 Sloan Research Fellowship, the 2005 NFS CAREER Award and the 2004 Best Paper Award Honorable Mention (with Prof. Yi Ma) at the European Conference on Computer Vision.

Rachel Ward
University of Texas at Austin, *rward@math.utexas.edu*
Department of Mathematics
PhD (2009), Computational Mathematics, Princeton University

Dr. Ward's research spans mathematical signal processing, applied harmonic analysis, compressed sensing, theoretical computer science, and machine learning. She is currently funded by an NSF CAREER award and an AFOSR Young Investigator Program Award.

Rebecca Willett
University of Wisconsin, *willett@discovery.wisc.edu*
Department of Electrical and Computer Engineering
PhD (2005), Electrical and Computer Engineering, Rice University

Her research interests include signal processing, machine learning, and large-scale data science. Specifically, inference from point process data, methods robust to missing data, high-dimensional data coupled with sparse and low-rank models, and streaming data.

Dr. Willett received the National Science Foundation CAREER Award in 2007, was a member of the DARPA Computer Science Study Group 2007-2011, and received an Air Force Office of Scientific Research Young Investigator Program award in 2010. She was a recipient of the National Science Foundation Graduate Research Fellowship, the Rice University Presidential Scholarship, the Society of Women Engineers Caterpillar Scholarship, and the Angier B. Duke Memorial Scholarship.

John N. Wright
Columbia University, *johnwright@ee.columbia.edu*
Department of Electrical Engineering
PhD (2009), Electrical Engineering, University of Illinois at Urbana-Champaign

Dr. Wright's research is in the area of high-dimensional data analysis. In particular, his recent research has focused on developing algorithms for robustly recovering structured signal representations from incomplete and corrupted observations, and applying them to practical problems in imaging and vision. His work has received an number of awards and honors, including the 2009 Lemelson-Illinois Prize for Innovation for his work on face recognition, the 2009 UIUC Martin Award for Excellence in Graduate Research, a 2008-2010 Microsoft Research Fellowship, and the 2012 COLT Best Paper Award (with Wang and Spielman).

Appendix II

Rapporteurs

Jordan Comins, Data Scientist
Virginia Tech Applied Research Corporation, Jordan.comins@vt-arc.org

Brian Hider, Project Manager
Virginia Tech Applied Research Corporation, brian.hider@vt-arc.org

Thomas Hussey, Senior Consultant
Virginia Tech Applied Research Corporation, twhussey@flash.net

Logistics & Planning

Brian Hider, Project Manager
Virginia Tech Applied Research Corporation, brian.hider@vt-arc.org

Katherine Peterson, Administrative Assistant
Duke University, kathy.peterson@duke.edu

Government Observers

David Han, *Office of the Assistant Secretary of Defense for Research and Engineering (Basic Science),* david.k.han.civ@mail.mil

Jiwei Lu, *Office of the Assistant Secretary of Defense for Research and Engineering (Basic Science),* jiwei.lu.civ@mail.mil

Workshop Chairs & Report Authors

Robert Calderbank, *Duke University*

Guillermo Sapiro, *Duke University*

Appendix III
Questions for Attendees

The two day workshop was organized around five plenary talks:

- Al Hero, "The need for new theory and new models"
- Rene Vidal, "Connecting Theory with Practice"
- Rob Nowak, "Integration of Sensing and Processing"
- John Wright, "Optimization"
- Yonina Eldar, "Historical Perspective on Sampling"

Each plenary was a survey talk, designed to stimulate small group discussions, each focused on a particular topic, with workshop participants rotating through the groups, and each group periodically reporting on findings to all participants. Initial discussion was free flowing, with subsequent discussion more tightly focused.

On the second day of the workshop, participants were asked to answer the following questions:

- What challenges should be addressed in the next 5, 10, and 15 years?
- What long-term research investments are necessary?