




Data-Driven Aerospace Engineering: Reframing the Industry with Machine Learning

Steven L. Brunton,^{*} J. Nathan Kutz,[†] Krithika Manohar,[‡] Aleksandr Y. Aravkin,[§] and Kristi Morgansen[¶]

University of Washington, Seattle, Washington 98195

and

Jennifer Klemisch,^{**} Nicholas Goebel,^{**} James Buttrick,^{††} Jeffrey Poskin,^{‡‡} 

Adriana W. Blom-Schieber,^{§§} Thomas Hogan,^{‡‡} and Darren McDonald^{**}

The Boeing Company, Seattle, Washington 98108

<https://doi.org/10.2514/1.J060131>

Data science, and machine learning in particular, is rapidly transforming the scientific and industrial landscapes. The aerospace industry is poised to capitalize on big data and machine learning, which excels at solving the types of multi-objective, constrained optimization problems that arise in aircraft design and manufacturing. Indeed,



Steven L. Brunton is an Associate Professor of Mechanical Engineering and a Data Science Fellow with the eScience Institute at the University of Washington. He received his B.S. in Mathematics from Caltech in 2006 and the Ph.D. in Mechanical and Aerospace Engineering from Princeton in 2012. He has co-authored three textbooks, and received the Army and Air Force Young Investigator Program (YIP) awards and the Presidential Early Career Award for Scientists and Engineers (PECASE). He is a Member of the AIAA.



J. Nathan Kutz received the B.S. degrees in physics and mathematics from the University of Washington, Seattle, WA, in 1990, and the Ph.D. degree in applied mathematics from Northwestern University, Evanston, IL, in 1994. He is currently a Professor of applied mathematics, adjunct professor of physics, mechanical engineering and electrical engineering, and a senior data science fellow with the eScience institute at the University of Washington.



Krithika Manohar is an Assistant Professor of Mechanical Engineering at the University of Washington. She received her B.S. in Mathematics & Computer Science from the University of Massachusetts Lowell (2013) and the Ph.D. in Applied Mathematics from the University of Washington (2018). She is a recipient of an NSF Mathematical Sciences Postdoctoral Research Fellowship and spent two years as a postdoctoral scholar and von Karman Instructor at Caltech.



Aleksandr Aravkin is an Associate Professor of Applied Mathematics at the University of Washington, and the Director of Math Sciences at the Institute for Health Metrics and Evaluation (IHME). Dr. Aravkin received B.S. degrees in Mathematics and Computer in 2004, an M.S. in Statistics and a Ph.D. in Mathematics in 2010, all from the University of Washington. He was a joint postdoctoral fellow in Earth and Ocean Sciences and Computer Science at the University of British Columbia from 2010–2012, and a research staff member at the IBM T.J. Watson Research Center from 2012–2015, as well as Adjunct Professor in Computer Science and IEOR at Columbia University.

Received 31 August 2020; revision received 3 December 2020; accepted for publication 10 December 2020; published online Open Access 16 July 2021. Copyright © 2021 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. All requests for copying and permission to reprint should be submitted to CCC at www.copyright.com; employ the eISSN 1533-385X to initiate your request. See also AIAA Rights and Permissions www.aiaa.org/randp.

^{*}James B. Morrison Professor, Mechanical Engineering. Member AIAA.

[†]Robert Bolles and Yasuko Endo Professor, Applied Mathematics.

[‡]Assistant Professor, Mechanical Engineering.

[§]Associate Professor, Applied Mathematics.

[¶]Professor and Chair, Aeronautics and Astronautics. Associate Fellow AIAA.

^{**}Boeing Test & Evaluation.

^{††}Boeing Commercial Aircraft Engineering.

^{‡‡}Boeing Research & Technology.

^{§§}Boeing Commercial Aircraft Engineering. Associate Fellow AIAA.

emerging methods in machine learning may be thought of as *data-driven* optimization techniques that are ideal for high-dimensional, nonconvex, and constrained, multi-objective optimization problems, and that improve with increasing volumes of data. This review will explore the opportunities and challenges of integrating data-driven science and engineering into the aerospace industry. Importantly, this paper will focus on the critical need for interpretable, generalizable, explainable, and certifiable machine learning techniques for safety-critical applications. This review will include a retrospective, an assessment of the current state-of-the-art, and a roadmap looking forward. Recent algorithmic and technological trends will be explored in the context of critical challenges in aerospace design, manufacturing, verification, validation, and services. In addition, this review will explore this landscape through several case studies in the aerospace industry. This document is the result of close collaboration between University of Washington and Boeing to summarize past efforts and outline future opportunities.

I. Introduction

DATA science is broadly redefining the state-of-the-art in engineering practice and what is possible across the scientific, technological, and industrial landscapes. The big data era mirrors

the scientific computing revolution of the 1960s, which gave rise to transformative engineering paradigms and allowed for the accurate simulation of complex, engineered systems. Indeed, scientific computing enabled the prototyping of aircraft design through physics-based



Kristi A. Morgansen is Professor and Chair of the William E. Boeing Department of Aeronautics & Astronautics at the University of Washington. Her research interests focus on nonlinear systems where sensing and actuation are integrated, stability in switched systems with delay, and incorporation of operational constraints such as communication delays in control of multi-vehicle systems. She is co-founder and co-director of the University of Washington Space Policy and Research Center and is an Associate Fellow of the AIAA.



Jennifer Klemisch is a senior data science consultant with expertise in the aerospace industry, emotional intelligence and linguistics. Previously she held the position of Advanced Analytics Strategic Leader of Boeing Test and Evaluation. She received her undergrad in Mechanical Engineering from University of North Dakota, MS in Finance from Webster University and MS in Data Science from Northwestern University. She holds a patent for an oxygen analysis measurement system and a patent pending for emotionally intelligent robotic pilot.



Nicholas Goebel is an aerodynamic flight test engineer at The Boeing Company where he works on applying data science and machine learning to flight test data. He received a Bachelors in Aerospace Engineering and Mechanics from the University of Minnesota and an M.S. in Applied Mathematics from the University of Washington.



James N. Buttrick is a Technical Fellow at the Boeing Company. His focus for the past 36 years has been developing automated manufacturing machines for aircraft assembly and composite part fabrication. He received his M.S. in Mechanical Engineering from the University of Washington in 1984, and his B.S. in Marine Engineering from the US Merchant Marine Academy in 1979.



Jeffrey Poskin is a mathematician in Boeing Research & Technology's Applied Mathematics group. He has expertise in constrained data fitting, geometric modeling, and numerical optimization. He received his B.S. in Mathematics from the University of Kansas in 2011 and a Ph.D. in Mathematics from the University of Wisconsin in 2017.



Agnes (Adriana W.) Blom-Schieber is a Technical Fellow at Boeing Commercial Airplanes, and an Affiliate Professor at University of Washington's Mechanical Engineering department. She is specialized in fiber-reinforced composite design, analysis, manufacturing and optimization. Her work includes development of design tools, model-based engineering, and multi-disciplinary optimization. She is a member of the ThermoPlastic composites Research Center (TPRC) technical advisory board. She is an Associate Fellow of AIAA and past chair for the Pacific Northwest AIAA section.

emulators that resulted in substantial cost savings to aerospace manufacturers. The Boeing 777 was the first aircraft to have been designed completely from simulation without a mock-up. In a similar fashion, machine learning (ML) and artificial intelligence (AI) algorithms are ushering in one of the great technological developments of our generation [1–7]. The success of ML/AI has been undeniable in traditionally challenging fields, such as machine vision and natural language processing, fraud detection, and online recommender systems. Increasingly, new ML/AI opportunities are emerging in engineering disciplines where processes are governed by *physics* [8,9], such as materials science, robotics, and control. An overview of opportunities in data-intensive aerospace engineering is shown in Fig. 1.

Advances in data-driven science and engineering have been driven by the unprecedented confluence of 1) vast and increasing data; 2) advances in high-performance computation; 3) improvements to sensing technologies, data storage, and transfer; 4) scalable algorithms from statistics and applied mathematics; and 5) considerable investment by industry, leading to an abundance of open-source software and benchmark problems. Nowhere is the opportunity for data-driven advancement more exemplified than in the field of aerospace engineering, which is data rich and is already built on a constrained multi-objective optimization framework that is ideally suited for modern techniques in ML/AI. Each stage of modern aerospace manufacturing is data-intensive, including manufacturing, testing, and service. A Boeing 787 comprises 2.3 billion parts that are sourced from around the globe and assembled in an extremely complex and intricate manufacturing process, resulting in vast multimodal data from supply chain logs, video feeds in the factory, inspection data, and hand-written engineering notes. After assembly, a single flight test will collect data from 200,000 multimodal sensors, including asynchronous signals from digital and analog sensors, including strain, pressure, temperature, acceleration, and video. In service, the aircraft generates a wealth of real-time data, which is collected, transferred, and processed with 70 kilometers of wire and 18 million lines of code for the avionics and flight control systems alone. Thus, big data is presently a reality in modern aerospace engineering, and the field is ripe for advanced data analytics with ML.

The use of data for science and engineering is not new [10], and most key breakthroughs in the past decade have been fundamentally catalyzed by advances in data quality and quantity. However, with an unprecedented ability to collect and store data [11], we have entered a new era of *data-intensive* analysis, where hypotheses are now driven by data. This mode of data-driven discovery is often referred to as the *fourth paradigm* [12], which does not supplant, but instead complements the established modes of theoretical, experimental, and numerical inquiry. In fact, we again emphasize the strong parallel with the rise of computational science in the past half century, which did not replace, but instead augmented theory and experiments. Just as today computational proficiency is expected in the workforce, so will data science proficiency be expected in the future. Thus, the

data-intensive transformation in the aerospace industry can learn from the digital transformation over the past decades. Perhaps the largest change will be in how teams of researchers and engineers are formed with domain expertise and essential data science proficiency, along with changes in research and development cycles for industry.

The aerospace industry presents a number of unique opportunities and challenges for the integration of data-intensive analysis techniques and ML. The transformative impact of data science will be felt across the aerospace industry, including 1) in the factory (design for manufacturability, re-use and standardization, process control, safety, productivity, reproducibility, inspection, automation, drilling, shimming); 2) in testing and evaluation (streamlining testing, certification, anomaly detection, data-driven modeling); 3) in the aircraft (inspection, design and performance, materials and composites, maintenance, future product development); 4) in human-machine interactions (advanced design interfaces, interactive visualizations, natural language interactions); and 5) in the business (supply chain, sales, human resources, and marketing). Because of the exacting tolerances required in aerospace manufacturing, many of these high-level objectives are tightly coupled in a constrained multi-objective optimization. Traditionally, this optimization has been too large for any one group to oversee, and instead, individual components are optimized locally within acceptable ranges. However, unforeseen interactions often cause considerable redesign and program delays. In the worst-case scenario, accidents may occur. With improvements in end-to-end database management and interaction (data standardization, data governance, a growing data-aware culture, and system integration methods), it is becoming possible to create a *digital thread* of the entire design, manufacturing, and testing process, potentially delivering dramatic improvements to this design optimization process. Further, improvements in data-enabled models of the factory and the aircraft, the so-called *digital twin*, will allow for the accurate and efficient simulation of various scenarios. In addition to these operational improvements, advances in data-intensive analysis are also driving fundamental advances in aerospace critical fields such as fluid mechanics [13,14] and material science [15]. Importantly, data science works in concert with existing methods and workflows, allowing for transformative gains in predictive analytics and design insights gained directly from data. Figure 2 provides a schematic of this process.

Despite these tremendous potential gains, there are several challenges facing the integration of data science in the aerospace industry. Foremost, due to the safety-critical aspect of aerospace engineering, data-driven models must be certifiable and verifiable, must generalize beyond the training data, and must be both interpretable and explainable by humans. Further, collecting vast amounts of data may lead to a *data mortgage*, where simply collecting and maintaining this data comes at a prohibitively high cost, and the data is cumbersome for downstream analysis. In the aerospace industry, customers are extremely diverse, with training and certification often



Thomas Hogan is an Associate Technical Fellow at the Boeing Company, where he has worked on mathematical modeling for nearly 20 years. He received a Bachelors degree in Mathematics & Computer Science from the University of Michigan Dearborn and his Ph.D. in Mathematics from the University of Wisconsin Madison.



Darren McDonald is a Technical Fellow at the Boeing Company, and he is a Fellow with the Society of Flight Test Engineers, as well as a Senior Member of AIAA. He has worked for over 20 years in Flight Test Engineering, Stability & Control. He received a Bachelors degree in Aerospace Engineering from Embry-Riddle Aeronautical University, Prescott, AZ.

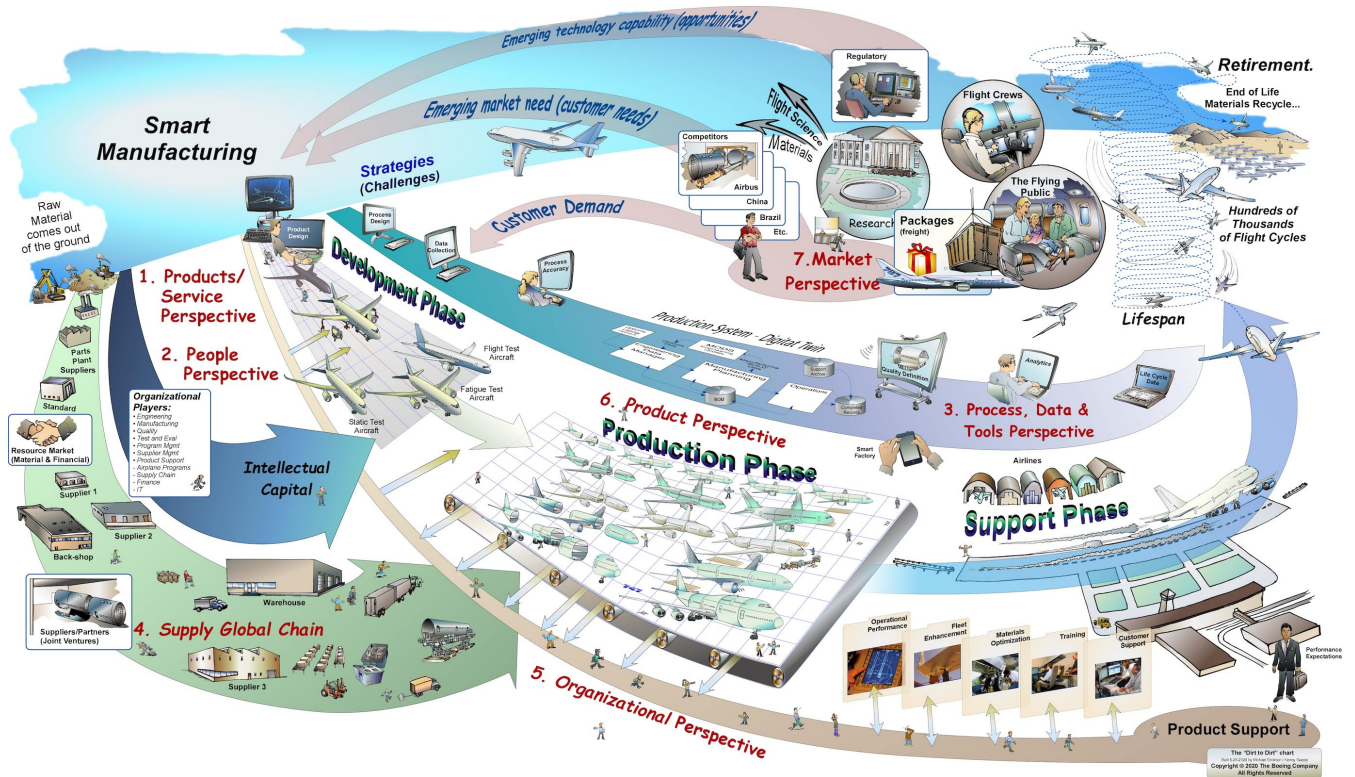


Fig. 1 Data science and machine learning have the potential to revolutionize the aerospace industry.

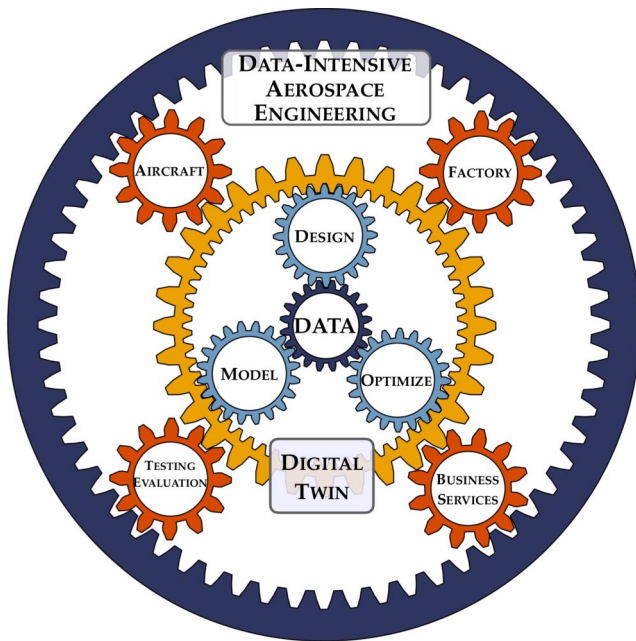


Fig. 2 Schematic overview of data-driven aerospace engineering.

done by a range of companies and regulatory authorities, respectively, motivating a level of robustness that is typically not required in other industries. There are also fundamental differences between commercial and defense sectors, limiting the sharing of designs, transfer of technology, and joint testing; in fact, at Boeing, testing and evaluation of commercial and defense vehicles have only recently been grouped under one organization, and these efforts still remain siloed.

The aerospace industry has always been a leader in optimization, and the earliest advances of the Wright brothers may be viewed as an optimization of the flight control system. In the century that followed,

much of the aerospace industry has been centered around a constrained, multi-objective optimization with an exceedingly large number of degrees of freedom and nonlinear interactions. ML algorithms are a growing set of data-intensive optimization and regression techniques that are ideal for these types of high-dimensional, nonlinear, nonconvex, and constrained optimizations. Aided by advances in hardware and algorithms, modern ML is poised to enable this optimization, allowing a much broader and integrated perspective. It is important to note that not all ML is *deep learning* or *artificial intelligence*. ML is simply optimization that is directed on data rather than first principles models, so that it fits naturally into existing efforts, but leverages a growing and diverse set of data.

In this paper, we will explore the evolving landscape of data-driven aerospace engineering. We will discuss emerging technology and how these are changing what is possible in aerospace design, manufacturing, testing, and services. These advances will be viewed through case studies, providing a review of past work and a roadmap for the future. Further, we wish to temper the excitement about the immense opportunities of data science with a realistic view of what is easy and what is hard. Finally, because the field is evolving rapidly, we want to establish a common terminology, taxonomy, and hierarchy for technology surrounding data science and its ML algorithms.

II. Machine Learning and Optimization

In this section, we will discuss a number of mathematical architectures that are central to the data-intensive modeling of aerospace systems. First, we will focus on understanding modern ML algorithms that will be critical to process data from the aerospace industry. We then provide a brief overview of applied optimization, which is the mathematical underpinning of ML. Next, we discuss specific ML extensions and considerations for systems governed by physics, and also discuss the importance of scalable and robust numerical algorithms. The treatment of these mathematical topics is far from comprehensive [9] and is instead intended to provide a brief, high-level overview. Potential uses for these algorithms will be explored in later sections.

A. Machine Learning

ML is a growing set of optimization and regression techniques to build models from data. There are a number of important dichotomies with which we may organize the variety of ML algorithms. Here, we will group these into *supervised* and *unsupervised* learning methods, based on the extent to which the training data is labeled. An approximate organization of these learning techniques, organized by task, is shown in Fig. 3. For example, in supervised approaches, a function ϕ mapping input data \mathbf{x} to outputs \mathbf{y} , also known as labels or targets, must be learned. Typically, the mapping may either predict the label

$$\hat{\mathbf{y}} = \phi(\mathbf{x}; \theta) \quad (1)$$

or model the joint probability distribution of the inputs and outputs

$$\phi(\mathbf{x}, \mathbf{y}; \theta) \quad (2)$$

In both cases, the models are parameterized by the parameters θ , which must be learned. There are numerous techniques to learn the structure and parameters of these mappings, including linear and nonlinear regression, genetic programming (i.e., symbolic regression), and neural networks (NNs). In addition to these dichotomies of ML, *reinforcement* learning, or *semisupervised* learning, provides yet another paradigm for learning where delayed reward structures can be achieved for tasks such as autonomy (self-driving cars, unmanned aerial vehicles, etc.).

There are four major stages in ML: 1) determining a high-level task or objective, 2) collecting and curating the training data, 3) identifying the model architecture and parameterization, and 4) choosing an optimization strategy to determine the parameters of the model from the data. Human intelligence is critical in each of these stages. Although considerable attention is typically given to the learning architecture, it is often the data collection and optimization stages that require the most time and resources. Indeed, exploring a broad and diverse set of ML architectures is a hallmark feature of today's ML-centric companies. It is also important to note that known physics (e.g., invariances, symmetries, conservation laws, constraints) may be incorporated in each of these stages. For example, rotational invariance is often incorporated by augmenting the training data with rotated copies, and translational invariance is often captured using convolutional NN architectures. In kernel-based techniques, such as Gaussian process regression and support vector machines, symmetries can be imposed by means of rotation-invariant,

translation-invariant, and symmetric covariance kernels. Additional physics and prior knowledge may be incorporated as additional loss functions or constraints in the optimization problem.

1. Supervised Learning

Supervised learning assumes that the training data \mathbf{x} has labels \mathbf{y} . If the labels are discrete, such as a categorical description of an image (e.g., dog vs cat), then the supervised learning task is a *classification*. If the labels are continuous, such as the lift profile for a particular airfoil shape, then the task is a *regression*. In the simplest form, the goal of supervised learning is to train a model to minimize the loss function

$$L = \|\mathbf{y} - \phi(\mathbf{x}; \theta)\| \quad (3)$$

where $\|\cdot\|$ is the root mean-squared error (RMSE).

Labeling the data with expert knowledge often makes it possible for supervised learning algorithms to automate complex tasks. Many of the dominant, industrially used algorithms are supervised in nature, including the ubiquitous methods of classification trees and support vector machines [3,9]. More recently, NNs have surpassed the performance of these methods, provided a sufficiently large data set is available to train the network [16]. In supervised learning, the availability of an extensive, high-quality labeled data set is crucial, as in the 2009 ImageNet data set [17], which enabled the training of a deep convolutional NN that outperformed all previous architectures [18].

2. Unsupervised Learning

Unsupervised learning, also known as data mining or pattern extraction, determines the underlying structure of a data set without labels. Again, if the data is to be grouped into distinct categories, then the task is clustering, whereas if the data has a continuous distribution, the task is an embedding. Unsupervised learning is an extremely challenging task as the algorithm is unguided by expert labels. Such algorithms are commonly used in an exploratory fashion to learn about data and the kinds of correlations that exist between features. The three most commonly used methods for unsupervised clustering include *k*-means, mixture models, and hierarchical clustering [3,9]. In each case, the number of distinct patterns in the data is usually specified by the user and refined as a tuned hyperparameter. Learning features in an unsupervised manner can often lead to future developments that are supervised.

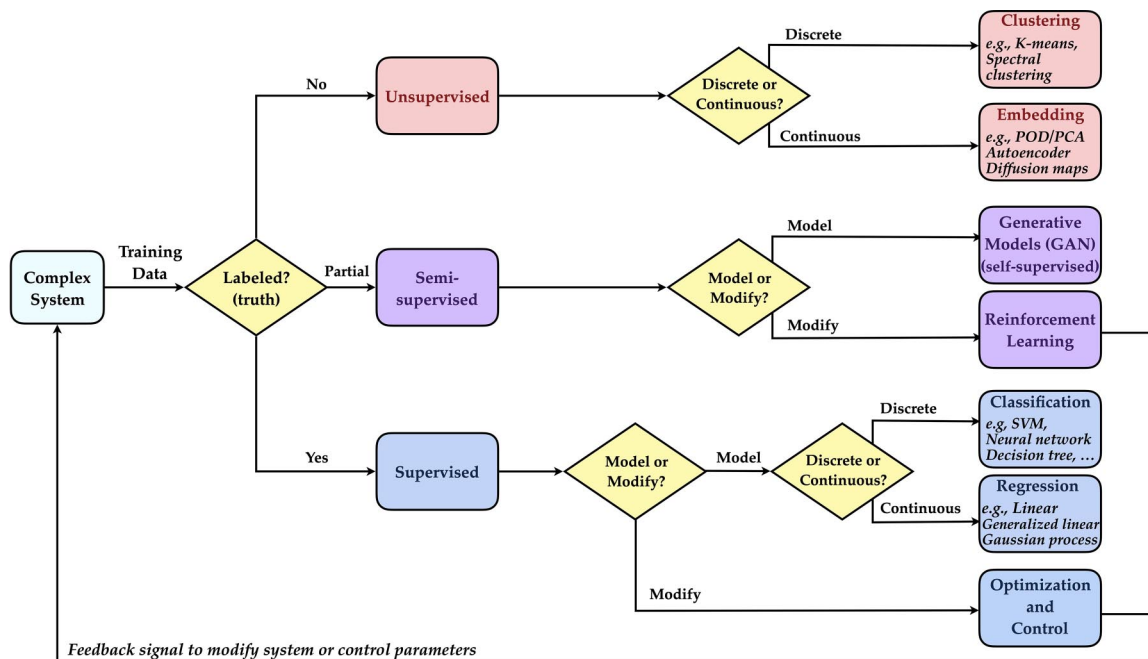


Fig. 3 Schematic overview of various machine learning techniques.

One of the standard approaches in embedding is to find a low-dimensional subspace or submanifold, parameterized by a latent variable z , that describes a high-dimensional state x . In this case, the goal is to find two functions, an encoder $z = \phi(x)$ and a decoder $\hat{x} = \psi(z)$, so that $\hat{x} = \psi(\phi(x)) \approx x$. The functions ϕ and ψ are implicitly parameterized by weights θ that must be tuned to minimize the following loss function:

$$L = \|x - \psi(\phi(x))\| \quad (4)$$

When the encoder and decoder are linear functions, given by matrices, then the optimal embedding recovers the classical singular value decomposition (SVD) or principal component decomposition (PCA) [9,14]. However, these functions may be nonlinear, defined by NNs, resulting in powerful autoencoders.

3. Reinforcement Learning

The power of reinforcement learning (RL) [19] lies in its ability to learn from interactions with the environment with goal-oriented objectives. Thus its application domain includes autonomy and control. This is unlike the two other dominant ML paradigms of supervised and unsupervised learning [9,16]. With a trial-and-error search, an RL agent senses the state of its environment and learns take appropriate actions to achieve optimal immediate or delayed rewards. Specifically, the RL agent arrives at different states s by performing actions a , with the selected actions leading to positive or negative rewards r for learning. Importantly, the RL agent is capable of learning delayed rewards, which is critical for many systems where a circuitous path to the optimal solution must be learned. Rewards may be thought of as sporadic and delayed labels, leading to RL often being classified as *semisupervised* learning. One canonical example is learning a set of moves, or a long-term strategy, to win a game of chess.

RL is often formulated as an optimization to determine the policy $\pi(s, a)$, which is a probability of taking action a given state s , to maximize the total reward across an episode. Given a policy π , it is possible to define a value function that quantifies the desirability of being in a given state:

$$V_{\pi}(s) = \mathbb{E} \left(\sum_t \gamma^t r_t | s_0 = s \right) \quad (5)$$

where \mathbb{E} is the expected reward over the time steps t , subject to a *discount rate* γ . Typically, it is assumed that the state evolves according to a Markov decision process, so that the probability of the system occurring in the current state is determined only by the previous state. Thus a large number of trials must be evaluated in order to determine an optimal policy. This is accomplished in chess and Go by self-play [20], which is exactly what many engineered systems are allowed to do to learn an optimal policy. Often, in modern deep RL, a deep NN is used to learn a *quality* function $Q(s, a)$ that jointly describes the desirability of a given state/action pair.

4. Deep Learning

Deep learning, or learning based on NNs with a deep multilayer structure, is often synonymous with ML and is the core architecture for many modern supervised and RL paradigms. NNs are particularly powerful due to their expressive representations of data and their diverse architectures [16]. The unparalleled success of these algorithms in ML is due to the availability of sufficiently vast and rich training data and modern computational hardware, which have enabled the training of exceedingly large NNs with millions or billions of free parameters.

Figure 4 shows a number of the leading architectures used in practice today in a variety of applications, including a simple feedforward architecture, a deep autoencoder that capitalizes on low-dimensional structure in data, and a deep convolution NN (DCNN) that is widely used to extract features for machine vision. The term “deep” refers to the number of NN layers (typically 7–10) used to map from inputs to outputs. NNs are universal function approximators [21], and they assume a compositional structure

$$y = \phi_1(\phi_2(\dots(\phi_n(x; \theta_n); \dots); \theta_2); \theta_1) \quad (6)$$

The flexibility of this compositional structure enables the construction of classification or regression maps between input and output data. With sufficient data, one can optimize for the NN weight

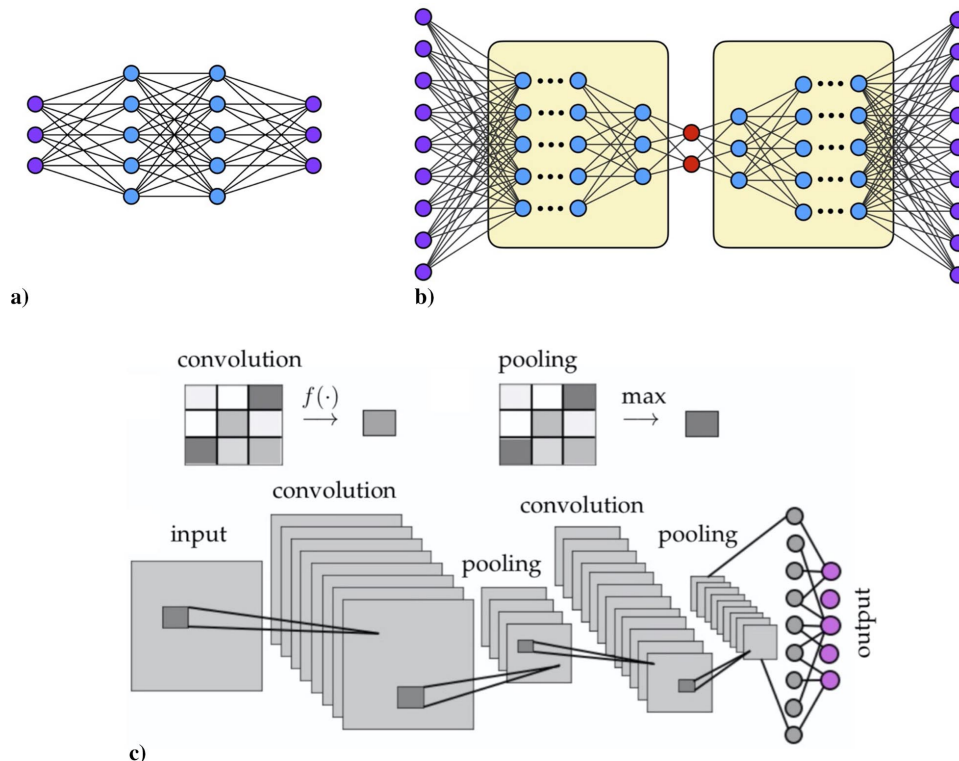


Fig. 4 Mathematical architectures of commonly used neural networks, including a) a feed-forward neural network, b) a deep autoencoder network, and c) a deep convolutional neural network.

parameters θ_n , usually via some form of stochastic gradient descent. Deep learning is commonly used in commercial settings, with DCNNs being the state-of-the-art for characterizing images and spatial correlations, and recurrent NNs (RNNs) enabling powerful speech and text recognition software. NNs are typically supervised learners and require a significant amount of data. They are also known to overfit to data and fail to generalize to new parameter regimes. Regardless, they are a powerful technology that can power many of the supervised learning tasks required in modern data science applications.

5. Physics Informed Learning

Physics informed learning [14,22–32] is of growing importance for scientific and engineering problems. *Physics informed* simply refers to our ability to constrain the learning process by physical and/or engineering principles. For instance, conservation of mass, momentum, or energy can be imposed in the learning process [22]. In the parlance of ML, the imposed constraints are referred to as regularizers. Thus, physics informed learning focuses on adding regularization to the learning process to impose or enforce physical priors. For the example of an NN model, this becomes

$$\underset{\theta_1, \theta_2, \dots, \theta_n}{\operatorname{argmin}} y = \phi_1(\phi_2(\dots(\phi_n(x; \theta_n); \dots); \theta_2); \theta_1) + \lambda g(\theta_1, \theta_2, \dots, \theta_n) \quad (7)$$

where the regularization $g(\cdot)$ imposes the desired physical constraint. The parameter λ is a hyperparameter allowing the user to impose an increasingly strong regularization to enforce this constraint. The importance of this constraint, or potentially multiple constraints, cannot be overstated in engineering and physics systems. Specifically, this is where known physics or physical constraints can be explicitly incorporated into the data-driven modeling process. Physics informed learning, often enacted with deep learning architectures, represents the state-of-the-art in ML methods for the engineering and physical sciences.

Rather than imposing physical constraints explicitly, an alternative is to learn embeddings based on physical models. This physics-guided paradigm involves learning embeddings from data produced by known first principles models of physics. The computation of these embeddings, in the context of aerospace and fluid dynamics, is often known as *modal analysis*, and has become increasingly data-driven (from either simulation or observation) in recent years [14,33,34]. The physical coupling between fluids and aerospace structures is particularly important, and the modes of these coupled interactions are impossible to discern by analyzing the Navier–Stokes equations and structural models alone; instead, they are determined by the boundary interactions between the coupled models. Thus, this physics guided architecture greatly enhances the understanding and design of robust engineering systems that can withstand complex interactions, turbulence, and instabilities. Improved reduced-order models (ROMs) of fluid dynamics may further aid efforts in flow estimation and control [35–41], reducing jet noise [42], and in improved turbulence models [13,43–50]. Modern modal analysis techniques, such as proper orthogonal decomposition (POD)/PCA [9], dynamic mode decomposition [51–53], Koopman mode decomposition [54,55], and resolvent mode analysis [56], naturally fall under the umbrella of unsupervised learning. Furthermore, the dimensionality reduction achieved by these methods enables low-latency, efficient downstream tasks such as system identification [57], airfoil shape optimization and uncertainty quantification [58], and reduced-order modeling [34,59–67].

B. Optimization

All of ML relies on optimization [9,14]. In fact, ML may be viewed as a growing set of applied optimization algorithms to build models from data. Mathematical optimization [68] comprises three interconnected areas: theoretical underpinnings, algorithm design and implementation, and modeling. Modeling allows us to bring in deterministic or physical descriptions of the real world, along with

information about structure of unknowns and statistical measures that describe beliefs about error and uncertainty. The scope of problems that can be formulated in this way is broad, ranging from sensing, estimation, and control [69–76] to ML [9,16,77] and decisions under uncertainty [78]. In all cases a modeling process gives rise to an optimization problem, where minimizing and maximizing over parameters leads to the desired inference or learning machine. Theoretical developments, such as convex and variational analysis, capture properties of problem formulations such as smoothness, convexity, and well-posedness of the problems themselves. In addition, deriving provable behavior of such algorithms is often important to guarantee performance.

A major distinction in optimization is between *convex* and *nonconvex* problems. Here, convexity refers to the property of the objective function to be minimized or maximized and the set of values being optimized over. Convex optimization problems [68] are extremely well studied, as there are fast and scalable generic solution techniques with performance guarantees. Convex objective functions will have a single global minima or maxima (i.e., the function has a single *hill* or *valley*), and gradient-based methods may be used to converge to this extremum. In contrast, nonconvex models, which have many local minima and maxima (i.e., the objective function has many local hills and valleys), form a much wider, and hence richer, class of problems. But, as such, there are no general scalable solution techniques or strong convergence guarantees.

1. Stochastic Algorithms

Many problems in ML involve training sets with millions of data points, making standard gradient computations prohibitively expensive. Stochastic optimization algorithms use random sampling to scale gradient descent algorithms to such data sets by using small subsets of the data at any given time. While these methods have a long history [79], recent developments have focused on extending these ideas to ML [80,81]. Recent theoretical breakthroughs have shown that the simplest algorithms that directly use sampled gradients, and even approximate gradients (subgradients) in the nonsmooth case, are provably convergent for a far wider problem class than was previously known [82]. These results justify the prolific use of these methods for general large-scale problems, such as training NNs. In addition, stochastic algorithms for structured nonconvex and nonsmooth problems make heavy use of the proximity operator [83,84], and can converge more rapidly than classic stochastic methods by exploiting variance reduction techniques to obtain improved search directions.

2. Role of Structure in Algorithm Design

A key theme in the field of optimization is to find the right problem classes that balance general applicability with specificity for the design of efficient algorithms. Rather than thinking of problems in generality such as convex versus nonconvex, more specific assumptions, reminiscent of physics-informed learning constraints, allow faster algorithms and better guarantees. Within convex optimization, the piecewise linear-quadratic class [85] captures a wide range of models and admits specialized solution techniques. Outside of convex optimization, the class of convex-composite [86] problems is a key generalization that has seen recent algorithmic development and analysis [87]. Coupled problems with multiple parameters have been solved for numerous applications with variable projection techniques [88]. Nonconvex composite models and algorithms have also been developed, with many applications involving challenging data-generating mechanisms and sparse regularization [89].

3. Atomic Operations for Nonconvex Nonsmooth Functions

There have been significant recent advances in nonconvex, nonsmooth optimization where the objective functions are not differentiable, and where parameters are constrained, for example, by physical bounds. NNs and deep learning models have been a particularly influential driver of methods for large-scale nonconvex models. Complementing this direction, nonsmooth models frequently arise as a means of imposing structure on the solution, for example, sparsity. In the remainder of this section, we discuss some of

these advances in more detail with references to survey literature in optimization.

When optimizing a smooth function, algorithms fundamentally rely on gradients to implement second-order methods, such as Newton's method or the Gauss–Newton and quasi-Newton variants (see [90] for an overview). These gradient computations involve matrix-vector products and equation solves, which we will call *atomic operations*. Most common algorithms can be decomposed into such operations.

Another key operation is that of the *proximity operator*, which has a long history and a tremendous range of recent applications [91,92]. Given a function f , we define

$$\text{prox}_{\alpha f}(z) = \underset{x}{\text{argmin}} \frac{1}{2\alpha} \|x - z\|^2 + f(x) \quad (8)$$

In words, we minimize the sum of the function and a scaled quadratic around a base point z and return the minimizing value. When the function f is arbitrary, evaluating the operator may be difficult or impossible. However, under moderate assumptions that are satisfied for a wide range of applications, the proximity operator has a closed-form solution or a provably fast computational routine. Many algorithms for solving nonsmooth nonconvex problems use the proximity operator as a subroutine, and are provably convergent [93].

C. Scalable and Robust Algorithms

Despite the growing abundance of measurement data across the engineering sciences, systems are often fundamentally low dimensional, exhibiting a few dominant features that may be extracted using dimensionality reduction [9,33]. The existence of low-rank patterns facilitates efficient models and sparse sampling, as there are only a few important degrees of freedom that must be characterized, regardless of the ambient measurement dimension. Here we discuss some of the tremendous advances in the past decades developing robust and scalable numerical algorithms for big data applications.

1. Randomized Linear Algebra

Massive datasets pose a computational challenge for traditional algorithms, placing significant constraints on memory, processing power, and computational time. Recently, the powerful concept of randomization has been introduced as a strategy to ease the computational load while still achieving performance that is comparable with traditional matrix factorization techniques. The critical idea of probabilistic algorithms is to employ some degree of randomness in order to derive a smaller matrix from a high-dimensional data matrix. The smaller matrix is then used to compute the desired low-rank approximation. Such algorithms are shown to be computationally efficient for approximating matrices with low-rank structure. Of particular interest are randomized routines for the computation of the SVD, (robust) principal component analysis (PCA), and CUR decompositions [94].

Several probabilistic strategies have been proposed to find a good smaller matrix, and we refer the reader to the surveys [95–98] for an in-depth discussion, and theoretical results. In addition to computing the SVD [99,100] and PCA [101,102], it has been demonstrated that this probabilistic framework can also be used to compute the pivoted QR decomposition [103], the pivoted LU decomposition [104], the CANDECOMP/PARAFAC (CP) tensor decomposition [105], and the dynamic mode decomposition [106]. It also helps frame computationally tractable ROMs [107,108]. It should be noted that the tech giants, such as Google and Facebook, routinely use randomized algorithms to analyze their large data sets.

2. Robust Dimensionality Reduction

Robust statistical methods are essential for evaluating real-world data, as advocated by John W. Tukey in the early days of data science [10,109]. Many techniques in dimensionality reduction are based on least-squares regression, which is susceptible to outliers and corrupted data. PCA suffers from the same weakness, making it *fragile*

with respect to outliers. To address this sensitivity, Candès et al. [110] introduced a robust PCA (RPCA) that decomposes a data matrix \mathbf{X} into a low-rank matrix \mathbf{L} containing dominant coherent structures, and a sparse matrix \mathbf{S} containing outliers and corrupt data:

$$\mathbf{X} = \mathbf{L} + \mathbf{S} \quad (9)$$

The principal components of \mathbf{L} are *robust* to the outliers and corrupt data in \mathbf{S} . The low-rank matrix \mathbf{L} is decomposed via the SVD into $\mathbf{L} = \mathbf{\Phi} \mathbf{D} \mathbf{V}^T$, where coherent features are given by the matrix $\mathbf{\Phi}$. We generally use the first r dominant columns of $\mathbf{\Phi}$, corresponding to the r features that explain the most variance in the data. The SVD provides the best rank- r least-squares approximation for a given rank r : $\hat{\mathbf{L}} = \mathbf{\Phi}_r \mathbf{D}_r \mathbf{V}_r^T$. The target rank r must be carefully chosen so that the selected features only include meaningful patterns and discard noise corresponding to small singular values. The subsequent left singular vectors $\mathbf{\Phi}_r$ are the desired low-rank features that span the columns of \mathbf{L} . The truncation parameter r may be determined using the optimal singular value truncation threshold of Gavish and Donoho [111].

RPCA has tremendous applicability for modern problems of interest, including video surveillance [112] (where the background objects appear in \mathbf{L} and foreground objects appear in \mathbf{S}), natural language processing [113], matrix completion, and face recognition [114]. Matrix completion may be thought of in terms of the Netflix prize, where a matrix of preferences is constructed, with rows corresponding to users and columns corresponding to movies. This matrix is sparse, as most users only rate a handful of movies. The goal is to accurately fill in missing matrix entries, revealing likely user ratings for movies they have not seen. We will demonstrate the use of RPCA for an aircraft shimming application in the case study in Sec. VIII.

III. Digital Twin and Enabling Technologies

Several key technologies are necessary to support the design, manufacturing, testing, and service of tomorrow's aerospace products, which will ultimately be enabled by a robust digital twin. These enabling technologies include sensors and the internet of things, a comprehensive digital thread, rapid data access and data storage, virtual reality to perform testing, ROMs and discrepancy models, uncertainty quantification, autonomy, and control. This section will review a number of these emerging technologies.

A. Digital Twins

Digital twin technology promises to revolutionize the entire manufacturing and engineering design landscape [115–119]. The goal of the digital twin is to bridge the physical and virtual worlds, providing a proxy environment to simulate, test, and evaluate model designs at a fraction of the cost of real-world implementation. The digital twin relies on an accurate, physics-based emulator that characterizes the statics or dynamics of a given system. Typically, this model will integrate a hierarchy of multiphysics and multifidelity models, which will be continually updated with data streams from the real world. To be more mathematically precise, many physics-based models will consist of a system of nonlinear partial differential equations describing the time-space evolution a system. Such an evolution equation can be represented as follows:

$$\mathbf{u}_t = \mathcal{N}(\mathbf{u}, \mathbf{u}_x, \mathbf{u}_{xx}, \dots, x, t; \boldsymbol{\beta}) \quad (10)$$

where \mathbf{u} is the system state, the subscripts denote partial differentiation, and $\mathcal{N}(\cdot)$ prescribes the generically nonlinear evolution. The vector $\boldsymbol{\beta}$ represents the parameters on which the system depends. Equation (10) also requires a set of initial and boundary conditions on a given domain. Typically, high-fidelity simulations of this system of equations are computationally involved, and it may be prohibitively expensive to simulate a truly multiscale system, such as the turbulent fluid flow over a full-scale wing at all scales. Instead, it is often necessary to leverage ROMs that capture dominant physical phenomena at a fraction of the cost. Of course, some systems are time independent and some are spatially independent. In either case, what

is critical is that a proxy, physics-based model exists that is capable of informing how a system behaves. From robots to manufacturing lines, an accurate virtual representation holds tremendous promise for technological advancement.

The success of digital twin technology centers on accurate virtual representations of the physical world. For precision manufacturing, for instance, current digital twin technologies do not provide the necessary level of fidelity for a number of processes. What distinguishes digital twins from traditional modeling efforts is the integration of multiphysics systems and components. Thus the digital twin often represents an entire engineering process versus individual components in the process, for which we may have good models. However, the digital twin must ensure end-to-end performance across the multiphysics system, placing stringent requirements on the fidelity of models and how they communicate. Discrepancy models (discussed below) provide an adaptive framework, whereby the digital twin can continuously learn updated, high-precision physics models over the course of time from its own sensor network. The integration of data and multiphysics models across a system is a grand challenge, requiring intelligent, robust, and adaptive architectures for learning and control. Many of the data-driven strategies discussed here are ideally poised to help in building accurate and viable digital twin models.

B. Sensor Technology and the Internet-of-Things

The aerospace industry generates tremendous volumes of data from a vast array of distributed sensors. With emerging *internet-of-things* sensing and communication capabilities, this volume of data will only increase. To avoid a data mortgage, where the majority of resources are spent collecting and curating data, it is critical that key features are automatically extracted and analyzed in real-time through edge computing. Thus, the paradigm of *big data* will shift to one of *smart data*. It is also important for algorithms to be robust to outliers. Outliers may correspond to sensor failures or saturations, although they may also signal important events that should be analyzed more carefully. Identifying where to place new sensors will also play a key role in improving efficiency and process control.

Many complex systems, such as a turbulent fluid or a large aerospace structure, have many degrees of freedom and are mathematically represented as a high-dimensional vector of data resulting from simulations or physical measurements. However, high-dimensional data often exhibit low-dimensional patterns, which is the foundation of dimensionality reduction and ML. The high-dimensional state $\mathbf{x} \in \mathbb{R}^n$ may then be efficiently represented in a low-dimensional basis $\Phi_r \in \mathbb{R}^{n \times r}$, for example, via RPCA above, so that $\mathbf{x} \approx \Phi_r \mathbf{a}$, where \mathbf{a} is a vector that approximates \mathbf{x} in terms of the first r principal components Φ_r . Often the structure of Φ_r is well-characterized from historical data (e.g., a library of human faces, or a set of point cloud scans of a particular aircraft part across several aircraft). In this case, the number of measurements required to estimate the full vector \mathbf{x} may be dramatically reduced from the ambient dimension n . These sparse sensors can be selected to best identify the coefficients \mathbf{a} in the basis Φ_r , thus enabling robust estimation of the high-dimensional state \mathbf{x} . There are a number of sparse sensing paradigms, including gappy sampling [120,121], empirical interpolation methods (EIM) [122–124], including the discrete empirical interpolation method, or *DEIM*, and compressed sensing [125–130].

Sparse sampling in a tailored basis [76,131,132] has been widely applied to problems in the imaging sciences as well as to develop ROMs of complex physics, such as unsteady fluid dynamics. Thus, even if we cannot measure the full state \mathbf{x} , it is often possible to estimate the state from $r \ll n$ point measurements in the *observation* vector $\mathbf{y} \in \mathbb{R}^r$ given by $\mathbf{y} = \mathbf{C}\mathbf{x} = \mathbf{C}\Phi_r \mathbf{a}$, where $\mathbf{C} \in \mathbb{R}^{r \times n}$ is the point measurement operator. Then, the reconstruction of the remaining state reduces to a least-squares estimation problem for the coefficients $\hat{\mathbf{a}} = (\mathbf{C}\Phi_r)^\dagger \mathbf{y}$, where \dagger is the Moore–Penrose pseudoinverse. This procedure was first used to reconstruct facial images from a subsampled pixel mask [120]. Importantly, it permits a drastic reduction in computation by solving for $r \ll n$ unknowns. The full state estimate $\hat{\mathbf{x}}$ is subsequently recovered using the feature basis:

$\hat{\mathbf{x}} = \Phi_r \hat{\mathbf{a}}$. The accuracy of reconstruction depends on the structure of the basis Φ_r and the choice of observations \mathbf{C} .

The number of observations can be greatly reduced by optimizing the sensors to maximize the accuracy of reconstruction. However, the combinatorial search over all $\binom{n}{r}$ possible sensor locations is computationally intractable even for moderately large n and r . However, there are convex relaxations that can be solved using standard optimization techniques and semidefinite programs in $\mathcal{O}(n^3)$ runtime. We advocate a greedy matrix volume maximization scheme using the pivoted matrix QR factorization [76,124]. This algorithm treats the selected measurements as optimal rows of the linear operator $\mathbf{C}\Phi_r$, which designs sampled features $\mathbf{C}\phi_i$ to be as orthogonal to each other as possible. Because the row selected columns ϕ_i are no longer orthonormal, greedy row selection methods attempt to maintain near-orthonormality of the features $\mathbf{C}\phi_i$ for a numerically well-conditioned inverse.

C. Reduced-Order Modeling

The engineering sciences increasingly rely on simulations as proxies for modeling expensive experimental systems. The complexity and dimension of these numerical simulations are growing rapidly due to increasing computational power and resolution in numerical discretization schemes. Many complex PDEs such as Eq. (10), which are critical for digital twin technologies, yield discretized systems of differential equations with millions or billions of degrees of freedom. These large systems, such as turbulent fluid flows, are extremely demanding, and may be prohibitively expensive, even for the most advanced supercomputers. Yet many dynamics of interest are known to be low-dimensional in nature, in contrast to the high-dimensional nature of scientific computing. ROMs help reduce the computational complexity required to solve large-scale engineering systems by approximating the dynamics with a low-dimensional surrogate [61–64,133–135].

To aid in computation, the selection of a set of optimal basis modes is critical, as it can greatly reduce the number of differential equations generated. Many solution techniques involve the solution of a linear system of size n , which generically involves $\mathcal{O}(n^3)$ operations. Thus, reducing n is of paramount importance. It is possible to approximate the state \mathbf{u} of the PDE using a Galerkin expansion:

$$\mathbf{u}(t) \approx \Phi_r \mathbf{a}(t) \quad (11)$$

where $\mathbf{a}(t) \in \mathbb{R}^r$ is the time-dependent coefficient vector and Φ_r is an optimal basis of orthogonal columns, typically generated by SVD; in this case $r \ll n$. We then substitute this modal expansion into Eq. (10) and leverage the orthogonality of Φ_r to yield the reduced evolution equations

$$\frac{d\mathbf{a}(t)}{dt} = \Phi_r^T \mathbf{L} \Phi_r \mathbf{a}(t) + \Phi_r^T \mathbf{N}(\Phi_r \mathbf{a}(t); \beta) \quad (12)$$

By solving this small system, the solution of the original high-dimensional system can be approximated. Of critical importance is evaluating the nonlinear terms in an efficient way, for example, using gappy POD or other sparse sampling techniques. Otherwise, evaluating the nonlinear terms still requires calculations of the high-dimensional function with the original dimension n . In certain cases, such as the quadratic nonlinearity of Navier–Stokes, the nonlinear terms can be computed once in an off-line manner. However, parameterized systems generally require repeated evaluation of the nonlinear terms as the basis may change with β . Regardless, ROMs allow one to approximate an n -dimensional system with an r -dimensional system, where $r \ll n$, making many computations possible that would otherwise be intractable. As such, ROMs are transforming high-performance computing by allowing for computational studies that have been intractable in the past. Recently, model discovery techniques are also providing low-order models of complex systems [22,24,57,136–138].

Despite the success of dimensionality reduction and reduced-order modeling across a range of problems, there are still many challenges

that remain. There are many systems, such as fluid turbulence, that defy a low-dimensional representation in a linear subspace as identified by SVD/PCA/POD. Indeed, turbulent flows often require hundreds or thousands of modes to describe the data, so that traditional projection-based model reduction approaches become infeasible. However, there is an opportunity to learn *nonlinear* submanifolds that accurately describe the data and are parameterized by fewer variables. Deep autoencoders provide one approach to learning such a nonlinear embedding where models may be identified [139,140]. Further, the related field of turbulence closure modeling is an area of active research that is advancing rapidly with techniques from ML [13,14,45,47].

D. Discrepancy Modeling

First principles modeling of physical systems has led to significant technological advances across all branches of science. For nonlinear systems, however, small modeling errors can lead to significant deviations from the true (measured) behavior. Even in mechanical systems, where the equations are assumed to be well-known, there are often model discrepancies corresponding to nonlinear friction, wind resistance, etc. Discovering models for these discrepancies remains an open challenge for many complex systems. There are many reasons why model discrepancies occur [141,142]. First, there may be measurement noise and exogenous disturbances. In this case, the Kalman filter may be thought of as a discrepancy model where the mismatch between a simplified model and observations is assumed to be a Gaussian process [69]. Second, the parameters of the system may be inaccurately modeled. Even worse, the structure of the model may not be correct, either because important terms are missing or erroneous terms are present. This is known as *model inadequacy* or model structure mismatch. Other challenges include incomplete measurements and latent variables, delays, and sensitive dependence on initial data in chaotic systems.

Discrepancy modeling centers on parameter and structural uncertainties. One can consider the following governing equations for a given engineering system given by Eq. (10). In general, there is also an output measurement,

$$\mathbf{y} = h(\mathbf{u}; \boldsymbol{\beta}) \quad (13)$$

from which the state may be estimated. The discrepancy modeling problem evaluates the difference between the model output of a quantity of interest $\boldsymbol{\phi}_m(t)$ and the observed value $\boldsymbol{\phi}_o(t)$:

$$\delta\boldsymbol{\phi}(t) = \boldsymbol{\phi}_o(t) - \boldsymbol{\phi}_m(t) \quad (14)$$

where $\delta\boldsymbol{\phi}$ is the discrepancy. The goal of discrepancy modeling is then to characterize the discrepancy $\delta\boldsymbol{\phi}(t)$, for example, using standard *Gaussian process regression* [143], dynamic mode decomposition [53] for approximating $\delta\boldsymbol{\phi}(t)$ with a best fit linear model, and/or model discovery for generating a nonlinear dynamic system [144].

E. Uncertainty Quantification

Model certification, credibility bounds, and other guarantees of performance are necessary for data-driven ROMs in the aerospace industry. Indeed, trustworthy ML is necessary for reduction to practice in almost any critical application area. The mathematical framework of *uncertainty quantification* (UQ) provides computational tools for evaluating probabilistic estimates of credibility and predictive capacity, and holds the key for bringing ML and AI into safety-critical domains. Without quantifying uncertainty in the model discovery procedure, one cannot provide estimates on the robustness and sensitivity of the models to observation error and model mismatch. In practice, this limits the applicability of all of ML for applications where quantifying the credibility of predictions is critical, such as human transportation systems with stringent safety regulations.

The mathematical architecture for UQ relies on a Bayesian perspective where predictions and quantification are given as probability distributions subject to a set of priors. Data-driven discovery must then be equipped with physically meaningful priors. It is also critical in developing a Bayesian scheme to separate dynamics from noise. Mathematically, model parameters $\boldsymbol{\beta}$ will be conditioned on the data \mathbf{Y} so that

$$p(\boldsymbol{\beta}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{Z}(\boldsymbol{\beta}))p(\boldsymbol{\beta}) \quad (15)$$

where $\mathbf{Z}(\boldsymbol{\beta})$ denotes the predictions provided by the candidate models, $p(\mathbf{Y}|\mathbf{Z})$ is the observation likelihood given by the observation model, and $p(\boldsymbol{\beta})$ is the prior on the model coefficients. A result of such UQ metrics is to have performance bounds and guarantees so that reduction to practice can be assessed and achieved.

F. Autonomy and Control

Modern aerospace systems, including manufacturing and operations, will rely on advanced autonomy and precision control. To date, there has not been an emergent and well-established paradigm on how to most effectively use large-scale data for autonomous control systems. Two modern grand challenge problems, robotics and self-driving cars, shown in Fig. 5, exemplify two very different paradigms. In the field of robotics, strong adherence to physics laws is imposed. Thus a robot is strongly constrained by our classical physics-based models for its movement and balance. However, critical use is made of high-fidelity sensors and Kalman filtering techniques, which jointly leverage models and data, in order to robustly control its sophisticated motions. In contrast, self-driving cars are largely physics-free. They are simply trained from exceptionally large data sets from sensors (vision, LIDAR, etc.), which attempt to integrate all possible scenarios the car may encounter in real-life driving scenarios. Both paradigms for leveraging data, physics-informed and physics-free, have experienced tremendous success in the past decade. They have also encountered fundamental limitations that must be overcome for the technologies to become commercially viable. Self-driving cars have been empowered by deep learning algorithms, which are known to have significant shortcomings in extrapolation tasks. Thus, some of the more spectacular

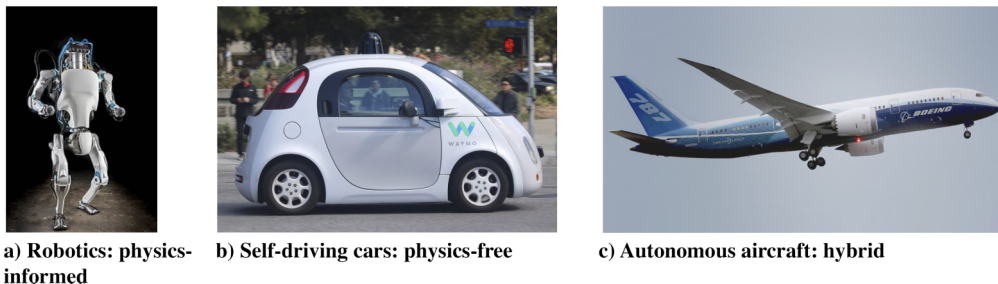


Fig. 5 Robotics and self-driving cars are both data-intensive examples of autonomy. Robotics is largely physics-informed, whereas self-driving cars are largely physics-free. Aviation requires both approaches. Panel (a) by Harrypotterrrrr, reproduced from https://commons.wikimedia.org/wiki/File:9_pic.png. Panel (b) by Grendelkhan, reproduced from https://commons.wikimedia.org/wiki/File:Waymo_self-driving_car_side_view.gk.jpg. Panel (c) by Dave Sizer, reproduced from https://commons.wikimedia.org/wiki/File:Boeing_787-8_maiden_flight_overhead_view.jpg.

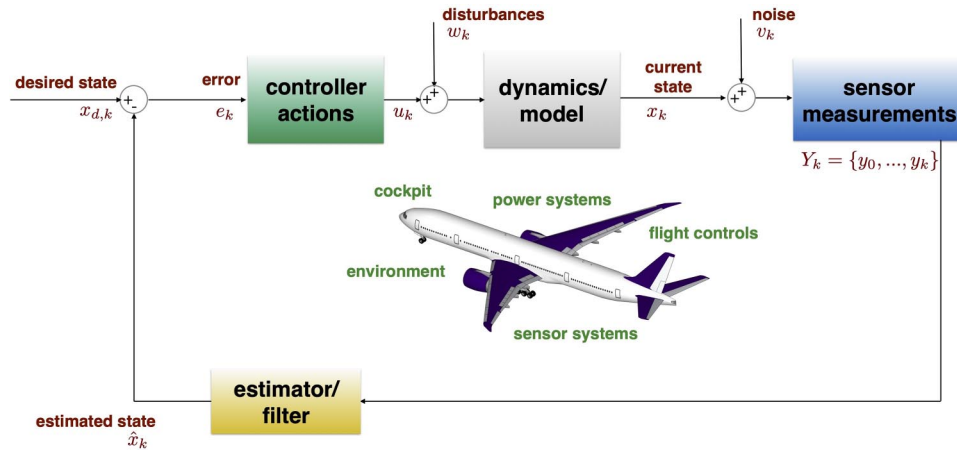


Fig. 6 Standard classical control feedback loop.

failings of self-driving cars have come from situations that were not part of their training sets. In contrast, our idealized physics-based models for complex and networked systems, such as robots, often fail to properly account for discrepancies between models and data, leading to control issues. The aerospace industry ultimately will need to leverage both paradigms in order to fully exploit their large-scale data sets.

A traditional feedback system involves guidance (what state do we want a system to have), navigation (what is the current state of the system), and control (how will we affect the system to achieve the desired state), shown in Fig. 6. Key elements of autonomy and control that must be considered when leveraging of data science include the following:

Performance criteria: Performance criteria for certification are generally specified in either the frequency domain or the time domain. In the time domain, the primary consideration is the transient response to a disturbance, relative to settling time, rise time, and overshoot. In the frequency domain, steady-state stability relative to harmonic forcing is considered (bandwidth, stability to loop closure, gain margin, phase margin).

Safety criticality: Certification generally requires demonstration of a system meeting design criteria to a certain probability of failure [e.g., “five nines” (1×10^{-5}), “seven nines” (1×10^{-7}), or “nine nines” (1×10^{-9})].

Human-in/on-the-loop: Effective control design must account for the interaction of the physical system with human pilots and passengers. For example, in pilot-induced oscillation, the reaction time of a human pilot relative to system dynamics can result in the pilot’s actions causing system oscillation to the point of instability. Modeling human interaction with a physical system is exceptionally complex.

Variable autonomy: When humans and autonomy interact, the relative amount of human authority versus autonomous authority must be selected. Typically systems are neither fully human operated nor fully autonomous. The greatest challenges appear when the level of autonomy changes during active operation.

IV. Aerospace Design

Aircraft design [145] starts with a set of requirements and ends with a production-ready vehicle design. Originally, the entire process took only a handful of engineers, from initial design, through multiple refinement and testing iterations, to final specifications. Modern aircraft design is much easier and would be impossible without many advancements that have occurred in chemistry, textiles, applied physics, cooking methods, fashion designing, and more.

Several innovations enabled production of two seminal aircraft in the early 1930s. The Boeing 247 and the DC-2 made commercial air travel viable. Their all-metal design, a major factor, depended on advances in structural design, new manufacturing ideas, and

development of new alloys. The design process involved significantly more wind-tunnel and flight testing, as well as new results in potential theory that enabled broader, more accurate aerodynamics calculations—calculations performed on slide rule. In fact, the aerospace industry has been a major driver in all of these fields due to its level of design complexity [146,147].

Advances in aerospace engineering, including jet propulsion, swept wings, radio navigation, supersonic travel, computerized control systems, and composite construction, enabled by subsequent innovations in commercial aircraft design, have each required more accurate calculations, greater computational power, and increasing levels of interdisciplinary interactions. Further advancements will require more innovations. As data management tools are developed and adopted to handle the increasing amounts of data required for each iteration, the cumulative knowledge obtained remains diffuse and disparate. And, although results in design exploration, surrogate modeling, and mixed-integer programming have enabled and advanced multidisciplinary design, handling significantly more design parameters, as seems imperative, will require optimization based on necessarily sparse exploration of increasingly high-dimensional design space. The aerospace industry is poised to both benefit from and drive advancements in data analytics, dimensionality reduction, data compression, and many more technologies broadly under the aegis of ML.

Aircraft design has also expanded to cover the entire life cycle, including data management, manufacturability, technical oversight, evaluation criteria, operations and operability, maintenance, and even disposal. It involves many competing objectives and constraints, including safety, environmental impact, and ergonomics. There is great potential from innovations in systems engineering, such as process control, robotics, flight scheduling, and flight path management. Within the design process, planning and scheduling could benefit greatly from existing algorithms in decision making and control, with and without uncertainty, which have proved challenging to incorporate.

A. Multidisciplinary Design Optimization

Aircraft design is, in large part, a constrained, multi-objective optimization problem [148–156]. Constraints include airplane-level requirements, such as range or fuel capacity, as well as production and business constraints, such as manufacturing costs or product cash flow. Design parameters may be discrete (e.g., engine count) or continuous (e.g., wing sweep); the objective is some measure of expected profit or quantities presumed correlated with profit. Although the level of success depends on the efficiency and accuracy of the models, analysis tools, and objective functions, it also depends on the dimensionality of the design space and the ability to explore it thoroughly enough to identify a superior design with confidence. Success also requires a robust design approach due to unavoidable uncertainties in predictions, algorithms, dynamics, etc.

Multidisciplinary design optimization (MDO) involves the use of numerical optimization in design when the constraints and

objective(s) depend on two or more analysis disciplines, for example, aerodynamics, structures, propulsion, controls, cost (including research/development, design, manufacturing, operation, end-of-life), performance (e.g., range, fuel burn), and environmental impact (e.g., noise, emissions). MDO can involve a diversity of parametric models that may include physical geometry, schematic layouts, or marketing mix. Regardless of model type, effective MDO requires complete automation of every step from parameter values to the objective function value for those parameters (or to the identification of the values as “infeasible”). This entails the models themselves, any processing required for their analysis, and postanalysis synergism.

ML is increasingly able to aid in this type of highly-constrained multi-objective optimization problem. It is possible to leverage the growing wealth of data to learn models for how the various parameters interact, providing fast and accurate surrogates for inverse design. For more explicit real-time optimizations, that is, control, ML models are currently being used to great effect in RL and model predictive control, which can natively balance multiple objectives and constraints.

The increased use of composite materials, in particular the complexity of fabrication methods, brings manufacturability to the forefront of disciplines considered in MDO studies. Reliable models are needed that can describe the output structure based on a variety of inputs, because the final structure of a composite material depends not only on the initial geometric design but also on the material and fabrication method used. Numerical modeling of fabrication methods, for example, hand lay-up, automated fiber placement, and vacuum forming, requires deep knowledge of the physical processes that govern the final structure [157]. Physics-informed ML may contribute to elevating first-principles-based models to the level of accuracy and efficiency required for an MDO study.

B. Model-Based Engineering

Model-based engineering (MBE) is an approach to product development and lifecycle management that focuses on using digital models and simulation to design, produce, maintain, and support products. Before the extensive use of digital artifacts across engineering domains, information and product specifications were transferred between engineering groups and product consumers through physical documents. Digital models improve reliability and robustness of the entire product lifecycle by ensuring that designers, producers, and consumers have access to evolving representations of physical assets. Techniques in data science are critical both in producing digital models that accurately capture asset behavior in the physical world and in standardizing digital model formats to ensure accessibility across the product lifecycle.

1. Design Diamond

The systems engineering design “V-model” depicts the major steps in the design and production of physical systems. The evolution of data generation in engineering processes as well as the technological development of robust digital models has led to a new view on the classical design V-model. In Fig. 7, the standard systems engineering design V-model is mirrored to create a design diamond that incorporates digital counterparts of a product at all stages of its lifecycle.

2. Digital Thread

The digital thread is a digital communications framework connecting authoritative sources of information from producers to consumers in standard formats throughout the lifecycle of a process, product, or system. It is the connection layer that ties together the digital models and physical processes guaranteeing that necessary groups are working with a consistent set of evolving information to support the product. One of the major challenges in implementing a digital thread is defining standard data formats that may interface with a wide variety of tools and processes. However, the presence of a digital thread in MBE is essential for high-fidelity digital twins.

3. Digital Twin

A digital twin is a virtual representation of the properties and behaviors of a specific instance of a physical system or process that enables prediction and optimization of performance and maintains synchronization with that physical system or process through its operational life.

The connection of a digital twin to its physical system is typically enabled through sensors gathering data in real-time during operations. A complicated system, such as a commercial aircraft, may consist of hundreds of connected subsystems each producing a wide variety of signals detailing performance characteristics. This leads to thousands of components producing diverse signals across varied communication channels, all in the course of real-time operations [158]. Identification of an appropriate subset of these signals when modeling a particular system process is a critical challenge in constructing a reliable and useful digital twin.

V. Aerospace Manufacturing

Manufacturing is a highly complex and dynamic process, involving the coordination and merging of several elaborate and precisely timed stages. In a modern manufacturing environment, tremendous volumes of data are being generated, stored, and analyzed to improve process quality, reliability, and efficiency. The data generated is inherently multimodal, including hand-written reports from humans

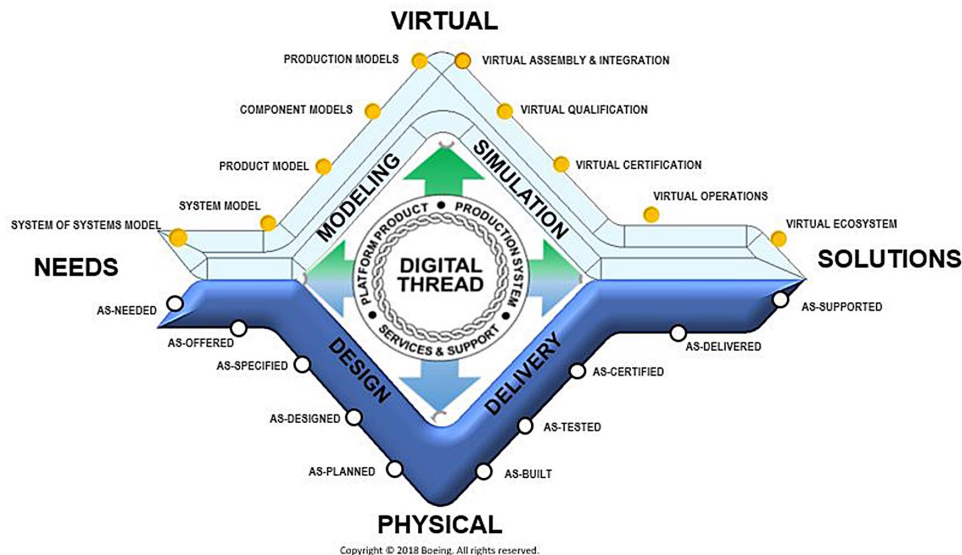


Fig. 7 Model-based engineering design diamond.

working alongside automation, high-fidelity data from metrology equipment, video feeds, supply chain logs, part catalogs, and results from in-process inspection, to name a few. There are several key opportunities to leverage ML and other data-intensive optimization techniques to improve manufacturing processes. Several areas of high-priority include part standardization; automation and robotics; streamlined assembly, including reduced measurements, processing, and inspection, toward *just-in-time* manufacturing; supply chain management; material design and fabrication; and nondestructive inspection. In this section, we will explore several of these areas.

A. Advanced Product Quality Planning

Advanced product quality planning (APQP) is a methodology for producing physical systems that are guaranteed to meet target requirements. It was developed in the late 20th century and has been used extensively throughout the automotive industry. Techniques in data science will greatly impact almost every technique and framework contained in APQP.

Data science and ML technologies have significant potential to improve performance in statistical process control. In aerospace manufacturing, this practice guarantees that parts are produced within a diverse set of specifications. Historically, process control has been implemented for manufacturing processes by extensive testing in labs before production and then tuning these processes throughout production efforts. Without consistent data storage and formats, this may lead to the inability to transfer past process control efforts onto new platforms and processes. Data science and ML may assist in transferring past efforts by defining standard data formats and producing robust digital simulation models.

The increased use of automation in aerospace manufacturing has also enabled new opportunities for real-time process monitoring. Manufacturing systems equipped with sensors gather real-time process data, which can then be used to train ML-based control models. Trained on past production data, these models may determine when a process will move out of specifications before possible human measurement and detection. Furthermore, the features extracted using ML methodologies can help determine ideal measurement and detection locations and can lead to significant reductions in labor and costs.

B. Standardization

The standardization of aircraft designs, manufacturing processes, parts, and machines is an open industrial challenge. New designs are often airplane-specific iterations that start the design process from scratch and ultimately offer no significant differences over prior versions. However, current manufacturing paradigms do not exploit or streamline designs to reuse and retool existing versions, resulting in inefficiencies and stresses on supply chains, maintenance, and the critical path of assembly. In aircraft design, some simple examples are brackets and fasteners. Brackets support wiring, hydraulics, systems, and fasteners for joining many types of structures and components. If data from prior designs can be extracted or harvested, then detailed comparisons can be performed and similarities can be identified, resulting in a reduction of unique parts that are airplane specific.

Data mining and ML are poised to revolutionize industry paradigms for standardization. Currently, modern digital tools for designing, inventory, and quality control produce a glut of data that often sits unused. This data can be mined for dominant features to inform future decisions and streamline the design process. To discover relevant, redundant patterns, feature selection criteria can be tailored to extract the desired topological, spatiotemporal, physical, and material properties. Recent strides have been made in interpretable ML models that directly incorporate sparsity-promoting, physical, temporal, and topological constraints into optimization objectives. The key features can be used to identify standard designs and reduce the number of unique designs or parts so that they can be shared across future designs and applications.

C. Automation

The aircraft industry has driven the development of automated machines to fabricate detailed parts, including machining metals,

composites fabrication, assembly of structure, painting, and inspection. These systems improve quality, reduce production cycle times and cost, and reduce repetitive injury to mechanics; however, they are expensive to develop, implement, and maintain. Due to the unique requirements of aircraft designs, materials, and tight tolerances, the decision to automate is complex and driven by specific manufacturing processes and applications. The development of successful automated systems starts with a focus on automating the process and understanding requirements. Thus, controlling the process is key to success. The tight tolerances of automation require monitoring and process control to ensure consistency, which in turn employs many sensors. Sensor feedback can provide insight into where processes are stable and where variation occurs. They tell us the level of control needed in specific area, process repeatability, what and where to measure, and a host of other outcomes.

When control is present, algorithms for ML control can be used to disambiguate the relevant process dynamics from control inputs. The tools of ML can also be used off-the-shelf on sensor data, not only to identify repetitive patterns and faults, but also pinpoint potential failures. Such failures can be predicted by studying the extent to which data deviate from dominant features, or by detecting outliers using robust feature extraction techniques on raw sensor data. The goal of robust feature extraction is to efficiently learn features and outliers simultaneously. One example is robust PCA, which has previously demonstrated success in predictive aircraft shimming applications [159] and is discussed in detail in Sec. VIII.

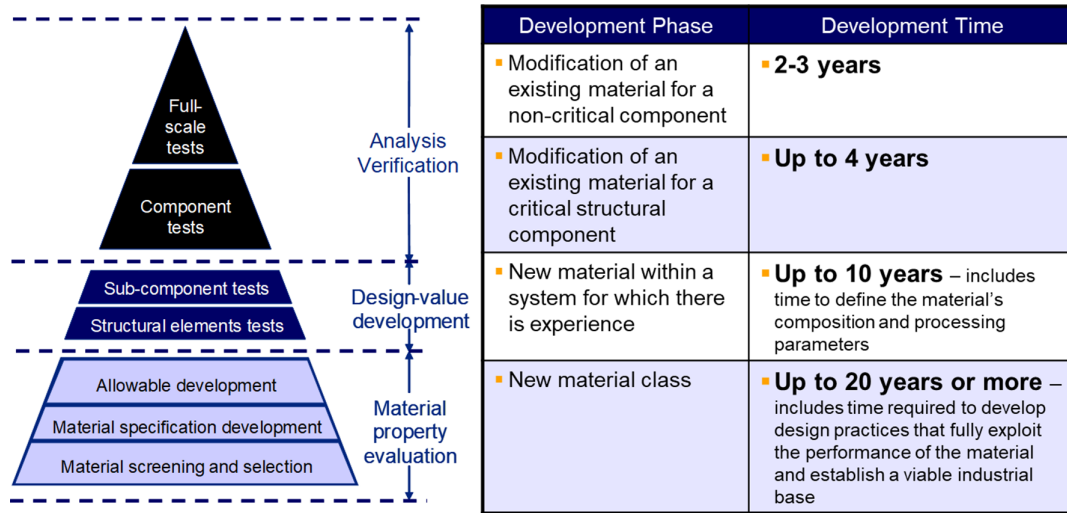
D. Assembly

Aircraft are manufactured in many pieces, including fuselage barrel sections, wings, stabilizers, and fasteners, that are joined as sections or subassemblies and ultimately get integrated into the final product. Wings are built from subassemblies of fabricated wing skins, spars, flaps, and ailerons; likewise, fuselages consist of multiple barrel sections that are assembled separately and then joined. Wings are joined to fuselages in a manufacturing position called “wing-to-body join.” The alignment, positioning, and joining of these assemblies require high precision tooling and part positioning systems. In some cases, automated positioning tooling controlled by metrology systems are used to align and fit structure. In addition, even though large sections are made to very tight tolerances, the combined tolerances result in very small gaps that must be shimmed to meet engineering requirements. Parts are measured with metrology equipment, three-dimensional models are generated to determine gaps between the mating parts, and automated machines are used to fabricate the shims that fill the gaps to ensure proper fit.

Throughout the process, enormous quantities of data are generated. Many sensors are used to provide feedback to control the automation, measure parts, and validate quality. The use of these data is expanding from building and assembling specific parts to being applied to predict future builds via ML and data analytics. Structural features mined from historical build data can be used to predict gaps in future builds, using only a targeted subset of measurements to infer the high-fidelity structure. This *sparse sensing* methodology was successfully applied for the predictive shimming of new wing-to-body join builds (Sec. VIII). Not only does sparse sensing bypass the heavy planning and processing required of high-fidelity metrology, but it also identifies a set of key structural features that may be analyzed for diagnostics and defects.

E. Materials

Adoption of new materials is a major challenge in the aerospace industry, due to the amount of structural testing done to certify a material. The standard approach is the building block shown in Fig. 8. This approach starts with a large amount of tests at small coupon levels, a much smaller number of tests at subcomponent scale, a handful of tests of subassemblies, and one or two full-scale test articles. Moreover, additional validation is needed when the manufacturing process or process parameters change, because they can introduce different features in the part. There are two areas where



SOURCE: R. Schafrik, GE Aircraft Engines, Technology Transition in Aerospace Industry, briefing presented at the Workshop on Accelerating Technology Transition, National Research Council, Washington DC, November 24, 2003

Fig. 8 Schematic materials building block diagram.

ML can have significant impact: physical testing and materials characterization.

The use of data analytics and ML, in addition to physics-based models, can result in a significant reduction in testing. First, we might be able to correlate behavior at larger scales to a small number of fundamental material properties by analyzing data on existing materials. That could already help filter out the most promising material candidates in the material screening and selection phase, without wasting time and money on extra characterization testing. In subsequent phases the sheer number of tests for new materials could be reduced by learning information from existing material systems. Secondly, in the absence of physical tests on new material parameters, physics-guided ML may be applied to process and structural simulations to correlate parameters to performance, thus enabling highly efficient exploration of material parameters. ML is also being used to design new materials [15], including superalloys [160–163].

More recently there has been an increased focus on material characterization for fiber-reinforced composites process modeling to reduce the amount of physical trial and error. Processing often happens when the material is uncured and at elevated temperature, requiring a different set of tests than is needed for the structural characterization. For example, numerical models used to predict the quality of thermoplastic composite laminates [164,165] require information on shear and bending behavior of the laminate, friction between layers (which changes when their relative orientation changes), and friction between the composite and any contacting tooling material (which in turn could have some surface coating). Most of these properties change with temperature, and some also depend on the forming speed and pressure. Here, sparse regression techniques in ML can help discover fundamental laws and relationships, resulting in significant reductions in the number of tests to be conducted, as well as the time and costs needed to introduce new materials.

F. Composite Fabrication

Composite part fabrication can consist of many steps, some of which are more data rich than others. Examples of the most time-consuming, defect-prone, or unpredictable processes are material laydown, forming, uncured part handling, in-process inspection, and autoclave curing. We will discuss some of these below.

Material laydown can be done manually, or by automated fiber placement (AFP) or automated tape laying (ATL). In AFP and ATL a robotic head places composite strips of material onto a tooling surface, building up each layer by making multiple passes to cover the surface. AFP is typically used for contoured parts, whereas ATL is most suitable for near-flat layup. Accuracy of the material placement

is important for the quality and final strength of the part and requires in-process inspection. Traditionally the in-process inspection has been a labor- and time-intensive part of the process. Vision and thermal inspection systems [166,167] are starting to replace manual inspection to measure material placement and detect other defects. The data produced by these systems are prime candidates for data analytics and ML [168], for diagnostics and defect detection.

A significant number of composite parts are not laid up in their final contour, but instead laid up flat and then mechanically formed. Examples are stringers and spars, which are formed from a flat laminate to an L-profile, hat-profile (see Fig. 9), or C-profile, in a machine and subsequently placed in a tool so they can be combined with a second profile or a skin to form an assembly once they are cured. These profiles often have to follow the contour of a fuselage or wing skin, which besides the global aerodynamic shape can have relatively aggressive geometric transitions when the skin thickness changes. The profile forming is often done in a machine, if the global contour is benign, while aggressively curved stringers are manually laid up ply-by-ply. Furthermore, placing the stringers in the tool is often done manually. The biggest risk with forming composite laminates is the generation of wrinkles when material is compressed locally to conform to its new shape. Small changes in pressure, material tack, or composition of the layup cause variability in the presence, location, and severity of the wrinkles. In addition, manual processes by definition have high variability, making it difficult to predict wrinkles for formed composites. Therefore every part is inspected and structural integrity is verified for every occurrence.

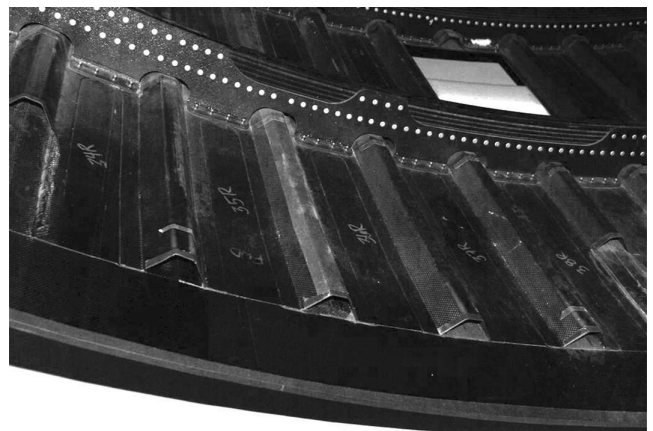


Fig. 9 Example of hat-shaped stringers on a 787 Fuselage [169].

and if necessary, repairs are made to recover any loss in performance. Data analytics and ML have the potential to improve composite forming processes in three critical areas: 1) support the development of physics-based models by identifying features that are dominant in the formation of wrinkles, 2) automate the detection and characterization of wrinkles in inspection data, and 3) improve part quality by using vision-based systems to track humans and determine patterns in their actions that result in better quality parts, which could lead to better work instructions or the development of an automated system that mimics the best manual process.

VI. Aerospace Verification and Validation

The testing phase of aircraft development, which includes extensive validation, verification, and certification processes, is poised to leverage data-driven methodologies and the associated wealth of predictive analytics. Testing a new product involves verifying function and performance guarantees and validating that the system meets certification and government regulations. A critical goal in modern data-driven flight testing is to improve aircraft safety and robust operation, while reducing time and expense of testing programs. This goal will be enabled by a greater understanding of the fully integrated product behavior, its robustness, and nuances, resulting in more effective models and digital twins. This knowledge will inform and improve upstream design and manufacturing processes and downstream services, ultimately enabling faster, more flexible aircraft design and customization by streamlining verification and validation cycles. Testing occurs at the component level, subsystem level, and the integrated system level, as depicted in the design diamond in Fig. 7. ML algorithms can be used to streamline the data collection and processing in each of these stages and to inform and refine earlier design and manufacturing stages. Other opportunities that are ripe for data-science enabled advances include the extraction and visualization of patterns that may be inaccessible to human analysis; sensor optimization, robust processing, and anomaly detection; identifying and characterizing discrepancies between models and physical devices; and using active learning to streamline the number of experiments and data required to validate models and manage uncertainty. In what follows, we frame the evolution of testing mainly through the lens of commercial aviation products, whose key technology trends are empowered by data digitization, digital-twins, and physics-based modeling to account for dynamic behaviors.

We first consider the type of data encountered in flight and lab tests. The amount of data generated by a flight test aircraft is diverse and vast, with upward of 200,000 multimodal sensor measurements during a single test. The measurements are a mix of airplane generated digital production measurements and analog flight test instrumented sensors. The multimodal sensors include strain gauges, pressure transducers, thermocouples, accelerometers, and video, among many others. Measurements are collected and stored asynchronously in time, with sampling rates varying from less than 1 Hz to upward of 65 kHz. Some measurement data is only collected when triggered or when a change is detected. All measurement data is stored for each flight test, generating gigabytes of data. Much of this data must be synthesized across sensor modalities and in time, both within a single flight test and across multiple tests. Similarly, high-rate asynchronous lab test data is collected in unique demanding environmental conditions, such as extreme vibration, icing conditions, sandblasting, extreme temperature, and high structural demands. These environments impose unique requirements on the sensors and can result in the sensors falling out of calibration.

The complexity of a typical commercial aircraft must also be considered. There are approximately 2.3 million parts on the Boeing 787 with 70 miles of wire and 18 million lines of source code in the avionics and flight control systems alone. Additionally, there are 10 major systems interacting together in a complex, dynamic, and rapidly changing environment. As for constraints, critical system components must have a failure rate of less than 10^{-9} or 10^{-12} , depending on the system, and backup components that seamlessly step in when required. Flight testing is the final validation step for every design and performance requirement. Certification is achieved

when each component, subsystem, and integrated system process can be shown to have a proven level of performance, repeatability, and accuracy. Because aircraft are designed to have near zero tolerance for failures, these failures are intentionally triggered during testing to evaluate the resulting effects on function and performance. The rarity of failures in the high volume of flight test data presents an additional challenge to ML algorithms, which require a sufficient number of examples to build models for failure.

A. Digitization

There is a significant opportunity to improve and automate verification and validation through data digitization, which will enable a wealth of downstream data analytics for pattern extraction and anomaly detection; see Fig. 10. To understand the benefits of a digitized workflow, it is helpful to understand the current testing landscape. Existing operations are rampant with mundane manual touch points, requiring significant process, time, and resources to validate that there are no quality lapses. Current workflows rely on experienced engineers to synthesize upward of 20 data sources to search for information to substantiate a report or troubleshoot a problem. Systematic efforts to leverage data across flights and programs and to perform exploratory analyses are virtually nonexistent.

A digitized flight test workflow that leverages ML would enable the automation of tasks that are currently laborious, reactive, last minute, discrete, and manual. The result would be a flow of information that is effortlessly available to all test stakeholders and participants from the preliminary concept to the final test report. Collection and aggregation of process metadata will enable visualization of core components, such as the test plan, test approvals, aircraft configuration, instrumentation configuration, and conformity status, with each component being updated as information becomes available. This centralized, real-time repository will provide engineers with the time and resources to process critical information, collaborate with peers, and engage with the test data as it is generated.

There are several significant challenges that face efforts to digitize verification and validation workflows. Digitizing the workflow cannot compromise the integrity of the information or its ease of access. Any changes must improve upon the users' ability to confidently assess the quality, reliability, and traceability of the data. However, the benefits of making operations more efficient by leveraging ML and uncertainty quantification make digitization worth the effort. It is expected that digitization will result in sweeping benefits across the production cycle and in-service life cycles.

B. Model-Based Validation

Historically, testing requirements were developed in component- or discipline-specific silos. These requirements were designed into test points as flight task cards to test specific configurations, or performance metrics. Critical or corner points of the performance envelope were identified as test conditions. Additional test points were determined by selecting evenly spaced points throughout the operating range of the component or aircraft. In conjunction with the test domain, the test procedures were aimed at isolating a small number of variables related to performance. The goal of test planning was to complete test conditions and test plans as efficiently as possible over the range of test parameters. It was typically the job of a test planner to consider readiness and dependencies of the test plans in order to identify the most efficient path through flight test.

Due to data capture and storage capacity limitations, data for flight tests were captured only for discrete single test conditions. Bandwidth and storage limitations prevented data from being used for real-time computing and recording. The phrase *data on* was used as the marker to begin a test condition where data would begin recording. As such, data for test conditions was gathered and stored in discrete segmented test condition increments. Only a small percentage of all data was ever stored and analyzed on any given flight. Further, if you weren't *data on* during a critical event in flight, either planned or unplanned, that data was lost without being recorded or analyzed. Engineers used paper strip charts that look similar to an EKG printout for real-time plotting. If the strip chart was not set to

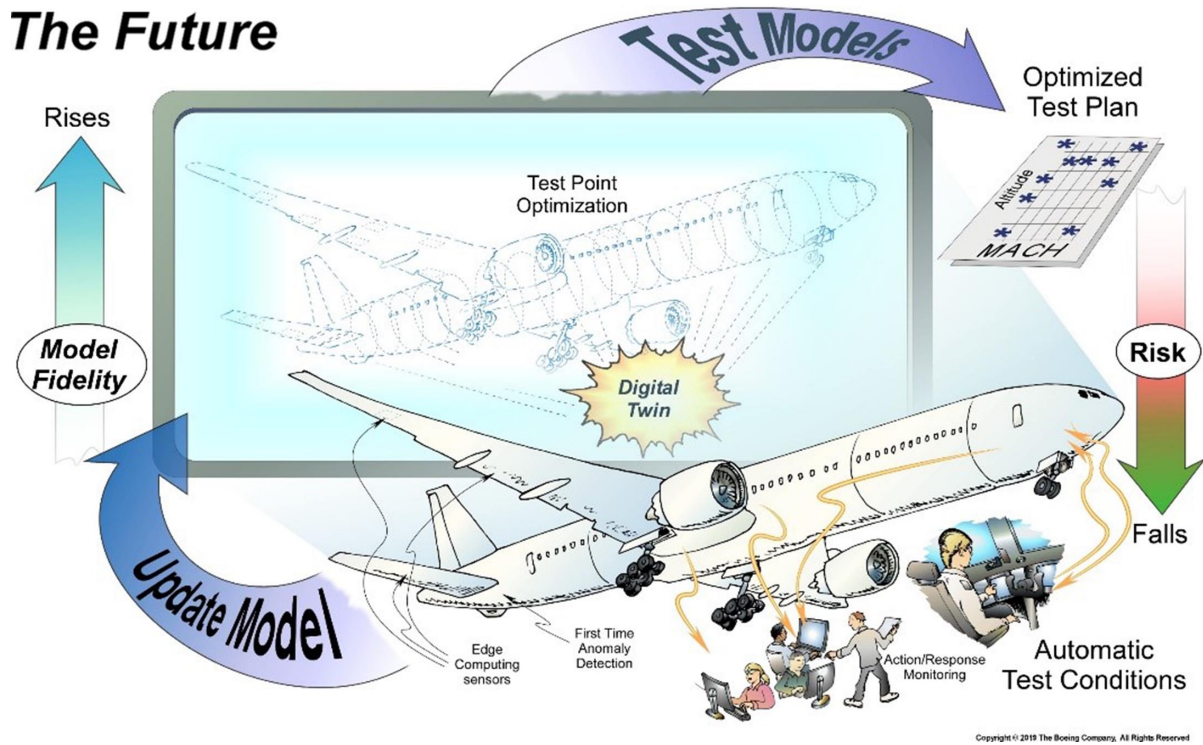


Fig. 10 Schematic overview of digital verification and validation processes, integrated with digital twin technology and optimized via machine learning.

plot the critical parameter during the test, there was no way to retrieve that data in flight. Cross plots were often created by hand with paper and pencil. Computers had limited computing capability and often only crucial calculations were computed in real-time. More intensive computing was performed by a mainframe computer after the test.

Technological innovations in computing and sensors have led to significant improvements in flight test data collection, storage, and visualization. Present day testing requirements are still largely developed in discipline-specific silos. Domain expertise has helped identify areas where tests can be modified to be more compatible or concurrent with other tests. Additionally, computational tools have been developed to aid human management and optimization of the flow of the test program. However, data science methods have yet to be used at significant scale to improve test optimization. Much of the improvement in present-day testing has been the result of improved computing hardware. Storage has improved so all data parameters can be recorded and stored for the duration of an entire flight. Data analysis has moved from mainframe computers to laptops. There have been some improvements in algorithms and calculation methods especially in the areas of data fitting and filtering. Data visualization has seen the greatest improvements: paper has been replaced with electronic strip charts and plots, and the use of colors and data triggers has allowed more data to be monitored and analyzed, with less cognitive overload. Although computational and visualization improvements have been significant, performance and evaluation metrics and algorithms have remained largely unchanged. Most data is still analyzed on a condition-by-condition basis, leaving large amounts of flight data unanalyzed.

Future testing will be heavily informed by the models that make up the digital twin. Safety-critical test conditions, where margins are thin, will be identified from models and simulations. Additional test conditions will be determined by a principled and data-informed selection, where physical tests will increase confidence in the model. Active learning [170,171] may be used to streamline this testing, reducing the number of physical queries required to validate a model. Effective test planning will involve a multidimensional and multi-objective optimization to efficiently achieve convergence of the coupled, multiscale, and multiphysics models. Test plans will need to be flexible enough to support model exploration and refinement when the model diverges from the physical world; that is, testing will

need to inform a discrepancy model for the system. The goal of flight test will evolve from completing as many test conditions as possible to increasing certainty in the multitude of physical, functional, and logical models that make up the digital twin. Such future testing will require a more seamless synchronization with the digital twin to the physical entity it represents, providing behavioral or functional predictions, which are then informed by measurement data. Robust and computationally tractable reduced-order modeling approaches, along with data assimilation and uncertainty quantification, will be required to validate and explore models on the physical asset in real-time. Data from these models will then be used to update the digital twins.

Flight test will likely continue to require discrete test conditions. However, an increased emphasis should be placed on analyzing data beyond what is required for the test condition. This holistic approach differs significantly from the past approach of only capturing data during the test condition to now making 100% of the data available for analysis. Modern techniques in ML will facilitate this analysis of high-dimensional, multimodal data, potentially extracting patterns and correlations that would have been inaccessible with previous approaches. For this model and data revolution to be successful, additional data from disparate sources will be required beyond what is needed to validate the digital twin. For example, meta data will link these data, including information about what parts were installed (configuration), what airport(s) were involved, and what were the qualitative characteristics of the aircraft. Combining this meta-data with the digital twin will be crucial to flight test of the future.

C. Leveraging Emerging Technologies and Algorithms

There are several other opportunities for enhanced flight test capabilities that are being enabled by rapidly advancing technologies and algorithms. As aerospace products become more advanced, it becomes increasingly difficult for human experts to manage the interactions between systems when troubleshooting a problem or designing a test with multiple failure conditions. However, ML and data visualization algorithms are enabling the analysis of vast, high-dimensional data sets, resulting in the identification of patterns and correlations that are intractable to human analysis. Identifying potential interactions, and ultimately, failure modes, is a combinatorially complex task; however, AI systems of the future may seamlessly

integrate decades of past test data to make informed decisions about likely complications. This advanced diagnosis and problem solving capability will only be possible with systemwide digital twin and digital thread efforts. In addition, considerable test time may be saved by automatically identifying sensor failure and detecting anomalous data [110,114,172], such as calibration drifts. Improved sensor fusion [173] and filtering algorithms, based on physics-informed ML models of the system and its components, will enable real-time diagnoses with confidence intervals. Sparse sensor optimization [76], or identifying which sensors to listen to for a given task, will also reduce the computational burden, enabling real-time decisions and updates to the flight plan, instead of postmortem analysis.

VII. Aerospace Services

In-service operations are the support provided to customers or by the customer operations team to support an operational fleet. From the initial delivery to long-term maintenance, there exist significant opportunities for leveraging data to improve and optimize maintenance and airside support. A major source of customer expense arises from aircraft damage, injury, delays to airside support, and unscheduled maintenance. The positive business impacts of autonomous airside operations and not incurring unscheduled maintenance can be met with emerging sensor and algorithmic innovations.

A. Autonomous Airside Support

Figure 11 shows the many data-driven aspects of ground service. Immediately upon arrival at an airport, preparations begin for receiving the airplane and preparing it for its departure. Critical tasks include 1) accurately positioning the aircraft in and out of the gate; 2) removing passengers, cargo, and waste; 3) reloading the aircraft with passengers, luggage, potable water, etc.; 4) servicing any aircraft fluids (oil, washer fluid, fuel) or maintenance items; 5) investigating any reported inbound issues; and 6) performing a visual pre-inspection check in the interior and exterior of the aircraft. Many of these tasks can be automated with modern technology, through robotic interactions, autonomous vehicles, or visual inspection using current computer vision algorithms. Importantly, modern IoT technology can integrate the diverse data streams in order to make an informed and accurate decision about the airside support tasks that require human intervention. Such data integration can provide significant savings, which allows human resources to be used in targeted and efficient ways. Importantly, even in this *fully* autonomous system, humans and

autonomy will need to interface in a safe, efficient, and effective manner. Consideration of robotic failure conditions and human activity awareness will be required.

B. Eliminating Unscheduled Maintenance

Unscheduled maintenance causes costly delays, wrecks havoc on a carefully optimized schedule, and inconveniences passengers, damaging the reputation of the airline carrier. Customers must balance the cost and time of preventative maintenance with the cost of unscheduled maintenance. Eliminating unscheduled maintenance would transform the industry, and this may be enabled by an expert aircraft system that identifies and communicates part anomalies and self-diagnoses wear, low fluids, software glitches, etc. This technical advance would require customer support centers to be able to download and process massive near-real-time data, run trend analysis, and make automated decisions in real-time. Current prescriptive models indicate that a particular part will fail within the next three flights with an 80% accuracy; although this information seems valuable, it is only marginally useful to improve operations. Increased value would be gained by understanding why the part is failing, or explaining the exact maintenance actions that need to be taken to rectify the situation. Knowing the right maintenance actions, for example, greasing the right bearing, updating software, or replacing the whole part, would have a significant impact on the efficiency and efficacy of maintaining the fleet. Being able to detect that a failure is about to happen, or has happened, is good. Being able to identify why the part is failing, and enabling a recommender system to assess the best course of action, is even better. Often, how the operator actually uses the system will violate the assumptions that were made when determining the maintenance schedules, which leads to early part failures. Leveraging part data would enable root cause identification and inform the design centers to enable future designs to be better aligned with in-service operations.

C. Closing the Loop with Design, Manufacturing, and Testing

When the digital thread connects data from aircraft design, manufacturing, and testing, it will be possible to leverage service data to improve each of these stages, and vice versa. The lifetime of an aircraft, from concept designs to retirement, spans decades, and the aging processes of new materials and manufacturing processes must be incorporated into the digital twin. These data-informed models will enable adjustments and refinements to the design and

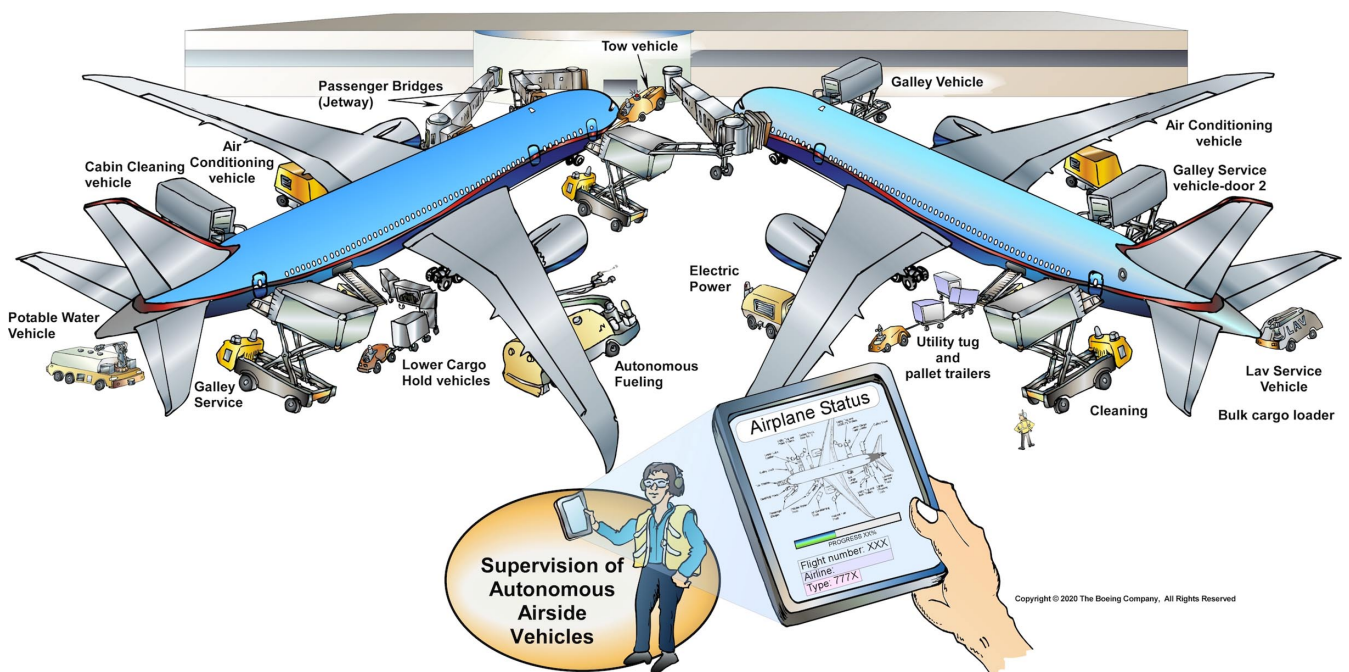


Fig. 11 Illustration of many potentially data-intensive aspects of ground service.

manufacturing procedures to improve the performance of future aircraft. Similarly, a more holistic digital twin, including models for aging and degradation, will be useful for maintaining and services fleets more effectively. These models will facilitate more accurate sensor filtering and data assimilation, as well as downstream tasks of anomaly diagnosis and detection.

VIII. Case Study: Predictive Assembly and Shimming

Aircraft are built to exceedingly high tolerances, with components sourced from around the globe. Even when parts are manufactured to specification, there may be significant gaps between structural components upon assembly due to manufacturing tolerances adding up across large structures. One of the most time-consuming and expensive efforts in part-to-part assembly is the shimming required to bring an aircraft into the engineering nominal shape. A modern aircraft may require on the order of thousands of custom shims to fill gaps between structural components in the airframe. These shims, whether liquid or solid, are necessary to eliminate gaps, maintain structural performance, and minimize pull-down forces required to bring the aircraft into engineering nominal configuration for peak aerodynamic efficiency.

Historically, parts have been dry-fit, gaps measured manually, and custom shims manufactured and inserted, often involving disassembly and reassembly. Recent advancements in three-dimensional scanning have enabled their use for surface measurement before assembly, known as *predictive shimming* [174–186]. Gap filling is a time-consuming process, involving either expensive by-hand inspection or computations on vast measurement data from increasingly sophisticated metrology equipment. In either case, this amounts to significant delays in production, with much of the time being spent in the critical path of the aircraft assembly.

In this case study, we present a recent strategy for predictive shimming [159], based on ML and sparse sensing to learn gap distributions from historical data and then design optimized sparse sensing strategies to streamline the collection and processing of data; see Fig. 12. This new approach is based on the assumption that

patterns exist in shim distributions across aircraft, and that these patterns may be mined from historical production data and used to reduce the burden of data collection and processing in future aircraft. Specifically, robust PCA (RPCA) [110] from Sec. II is used to extract low-dimensional patterns in the gap measurements while rejecting outliers. RPCA is based on the computationally efficient SVD [9,187] and yields the most correlated spatial structures in the aircraft measurements, identifying areas of high variance across different aircraft. Next, optimized sparse sensors [76,132,188] are obtained that are most informative about the dimensions of a new aircraft in these low-dimensional principal components. The success of the proposed approach, known within Boeing as PIXEL Identification Despite Uncertainty in Sensor Technology (PIXI-DUST), is demonstrated on historical production data from 54 representative Boeing commercial aircraft. This algorithm successfully predicts 99% of the shim gaps within the desired measurement tolerance using around 3% of the laser scan points that are typically required; all results are rigorously cross-validated.

This approach to predictive shimming combines robust dimensionality reduction and sparse sensor optimization algorithms to dramatically reduce the number of measurements required to shim a modern aircraft. In particular, RPCA from Sec. II is used to extract coherent patterns from historical aircraft production data. Thus, RPCA is used to develop low-dimensional representations for high-dimensional aircraft metrology data (e.g., laser scans or point cloud measurements). Shim scan data is collected across multiple aircraft, either leveraging historical data, or collecting data in a streaming fashion as each new aircraft is assembled. The shim measurements for a region of interest are flattened into column vectors $\mathbf{x}_k \in \mathbb{R}^n$, where n corresponds to the number of measurements and k refers to the aircraft line number. These vectors are then stacked as columns of a matrix $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_m]$, where the total number of aircraft m is assumed to be significantly smaller than the number of measurements per aircraft, that is, $m \ll n$.

Because of tight manufacturing tolerances and a high degree of reproducibility from part to part, it is assumed that the matrix \mathbf{X}

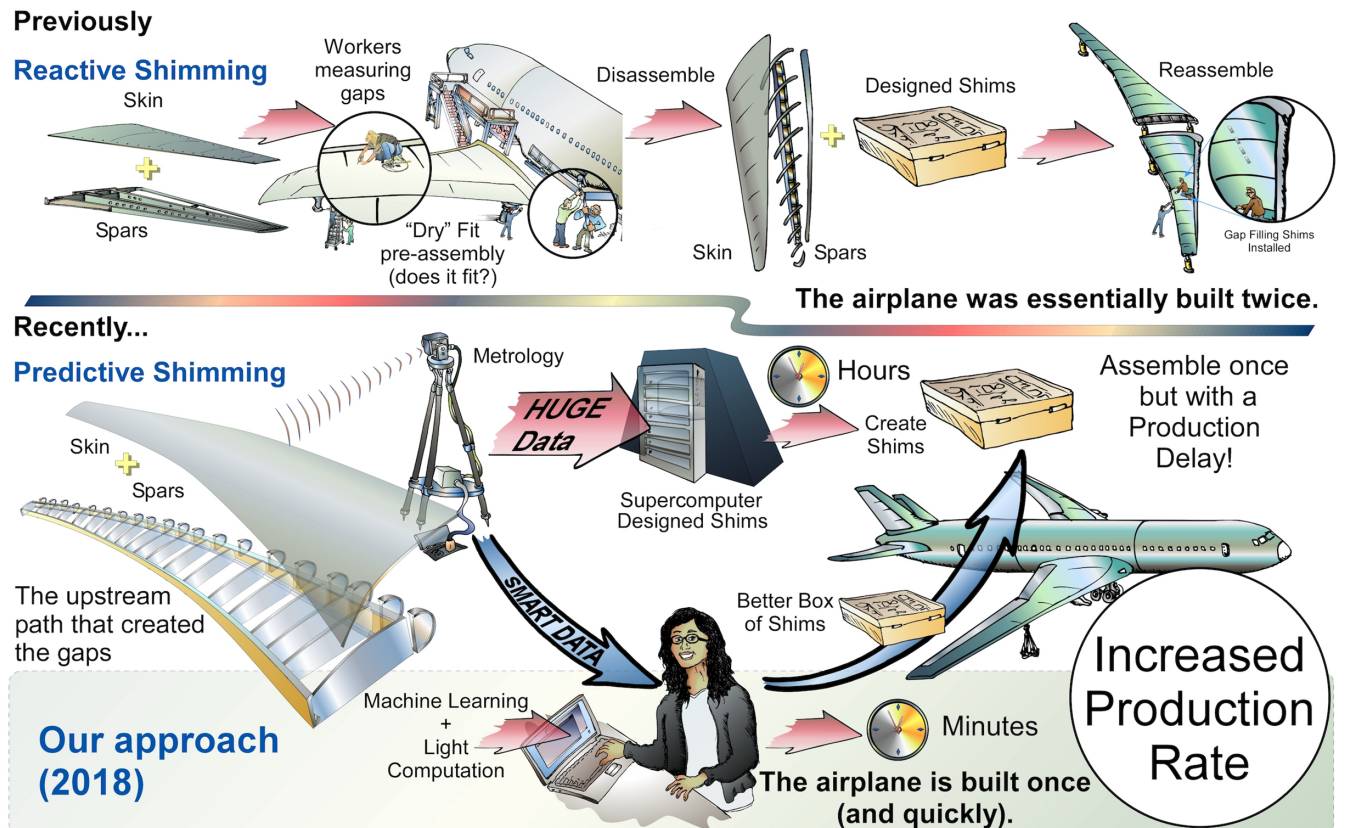


Fig. 12 Cartoon illustrating recent progress in predictive shimming in the last decade.

possesses low-rank structure. As described above, there may be sparse outliers that will corrupt these coherent features, motivating the use of RPCA to extract the dominant features. Next, sparse sensors are identified that maximally inform the patterns in future aircraft using sparse optimization techniques from Sec. II.B. The goal is to identify a small number of key locations that, if measured, will enable the shim geometry to be predicted at all other locations; this is possible because of the low-dimensional structure extracted through RPCA. Only measuring at these few locations will dramatically reduce measurement and computational times, improving efficiency.

We demonstrate the PIXI-DUST architecture on production data from a challenging part-to-part assembly on a Boeing aircraft. This data set consists of 10,076 laser gap measurements of the part assembly for 54 different production instances of the same aircraft type. Measurement locations are aligned between the instances, making the data amenable to SVD. We build a low-order model of the shim distribution using RPCA and then design optimized measurement locations based on these data-driven features. We train the model on 53 aircraft and then validate on the remaining aircraft; this process is repeated for all 54 possible training/test partitions to cross-validate the results. Thus, a data matrix $X \in \mathbb{R}^{10076 \times 53}$ of training data is constructed, in which each column contains all of the shim gaps for one aircraft, and each row contains the measured gap values at one specific location for all aircraft in the training set.

Figure 13 displays the seven separately manufactured shim segments, as well as the sensor ensembles for each shim; the error distributions are given in Fig. 14. Prediction results are shown in Table 1. Prediction accuracy is vastly improved, and 96–99% of the shim gap locations are predicted to within the desired machining tolerance. Furthermore, we note that the rates of optimal measurements r vary from anywhere between 2 and 6% of all points within the shim, which indicates that some shims are higher-dimensional and require more features (hence, sensors) to be fully characterized. This is also reflected in the sensor ensembles in Fig. 13.

This case study demonstrates the ability of data-driven sensor optimization to dramatically reduce the number of measurements required to accurately predict shim gaps in aircraft assembly. Sparse sensor optimization was combined with robust feature extraction and applied to historical Boeing production data from a representative aircraft. With around 3% of the original laser scan measurements, this learning algorithm is able to predict the vast majority of gap values to within the desired measurement tolerance, resulting in accurate shim prediction. These optimized measurements exhibit excellent cross-validated performance and may inform targeted, localized laser scans in the future. Reducing the burden of data acquisition and downstream computations has tremendous potential to improve the efficiency of predictive shimming applications, which are ubiquitous and often take place in the critical path of aircraft assembly. Thus, streamlining this process may result in billions of dollars of savings.

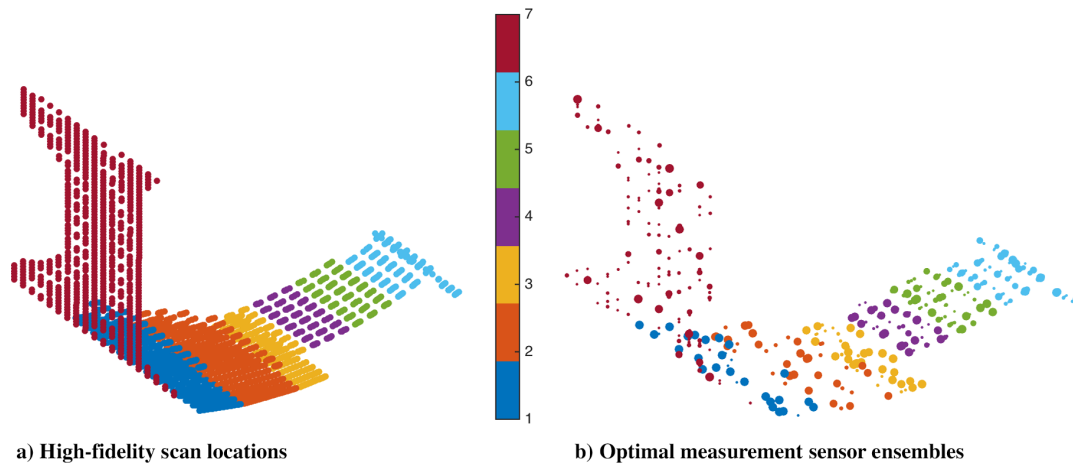


Fig. 13 Gap measurement locations segmented into seven shims (left) and ensembles of selected sparse sensors (right). Reproduced with permission from [159].

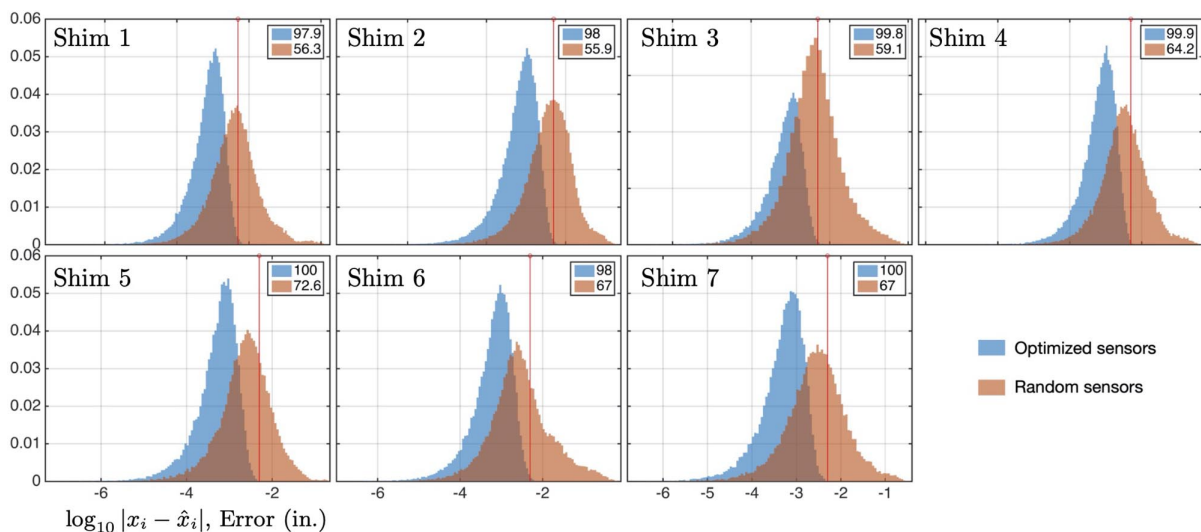


Fig. 14 Absolute error distribution for optimal (blue) vs random (orange) sensors on validation tests. The red line represents the desired measurement tolerance. Reproduced with permission from [159].

Table 1 Segmented prediction results show vastly improved prediction accuracies, with 97–99% of gaps predicted to within the desired 0.005-inch measurement tolerance

	Shim No.						
	1	2	3	4	5	6	7
Percent accurate	97.90	98.05	99.82	99.94	99.99	99.03	99.97
Optimal sensors (avg)	26	26	25	26	25	26	25
Total points	1003	1116	453	692	709	768	664

IX. Case Study: V-22 Osprey

The V-22, shown in Fig. 15, is a multirole tiltrotor combat aircraft, designed to take off and land like a helicopter, but with the range and speed of a fixed-wing aircraft. These impressive capabilities came with considerable engineering challenges, including the mechanics of the tiltrotor, the aerodynamics, and the control systems. The development and test of the V-22 were mired in delays, cost overruns, and safety mishaps. The purpose of this case study is to identify how the use of data science and a digital twin could have benefited the design and testing of a technologically advanced new aircraft, such as the V-22. The intent of this retrospective is not to critique the V-22, which is by all accounts an engineering marvel, or the design and testing program; many of the modern technologies discussed here did not exist or were in their infancy at the time. Instead, we present a summary, based on lessons learned from the V-22 program, where modern digital twin technology would have been beneficial to safety and facilitated an on-schedule design.

A 2001 review panel on the V-22 recommended the program “Extend high-rate-of-descent testing, formation flying (and other deferred flight tests as appropriate) to sufficiently define and understand the high-risk portion of the flight envelope under all appropriate flight conditions” [189]. The panel went further and recommended that the results of the high risk, high rate-of-descent testing be used to update operating limitations as well as the flight simulator used for crew training. The call for additional high rate-of-descent testing came as a result of a fatal accident in April 2000 involving both high-rate-of-descent and formation flying. The cause of the accident was attributed to an aerodynamic condition known as a vortex ring state (VRS), which is an unsteady aerodynamic condition that occurs for rotorcraft operating at low forward speed and high rate-of-descent. In this condition, the rotorcraft does not fly away from the rotor wake,

resulting in highly unsteady airflow. For a helicopter, this often results in a high rate-of-descent, and for a tiltrotor the result is typically roll-off. Both situations are perilous, especially in close proximity to the ground. VRS still remains a significant area of research.

The goal of a digital twin is to bridge the gap between the physical world and the virtual world. Two aspects of a digital twin that differentiate it from traditional modeling are 1) continuously learning and updating internal models with real-world data, and 2) integration of data from one or more multiphysics models across a system. For a digital twin to continuously learn, it needs to be tested, updated, and re-tested. In the case of the V-22, a digital twin may begin with a model of VRS such as the one presented in [189], which is merged with the model and flight test data for the specific aircraft. Discrepancy modeling can be used as a data-driven method to explain and update model divergence. On at least two occasions V-22 pilots experienced significant uncommanded roll events during formation flying [190]. As a result of this, and the aforementioned accident, the V-22 kicked off an extensive formation flying and high-rate-of-descent flight test campaign. Combining a VRS model with formation-flying models would be an ideal opportunity to integrate multiphysics models across a system. The goal of integration as data becomes available is to work toward model convergence, or the bridging of the physical and virtual such that the digital model accurately represents the physical environment. The integration of unsteady, nonlinear, multiphysics models is no easy task. Implementing a real-time reduced-ordered model may help discover and explore areas of significant model discrepancies in flight test. The use of physics informed learning and custom regularization in the ML optimizations may aid in maintaining physical constraints as the two models are merged to form a digital twin. With that enhanced model in hand, there is benefit in learning from a high number



Fig. 15 V-22 Osprey. Image by Peter Gronemann, reproduced from <https://en.wikipedia.org/wiki/File:V22-Osprey.jpg>.

of simulations. Flight testing could be reduced to the most critical test conditions: safety, mission critical, and model confirmation. Certainly, this simple case study is not without conjecture. However, the tools presented in this paper will form the basis for data-driven hypotheses that serve as the roadmap for future aerospace engineering possibilities.

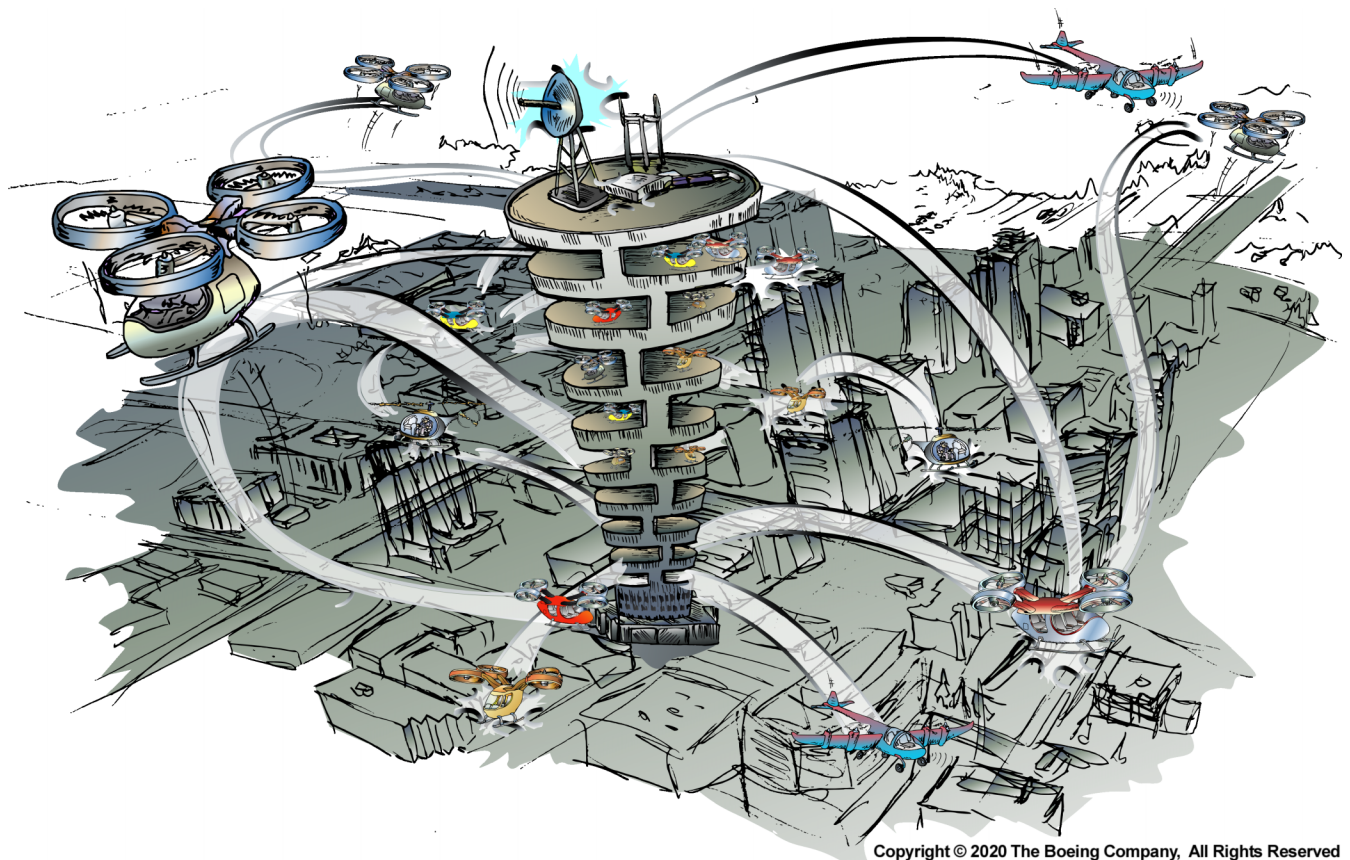
X. Case Study: Urban Air Mobility

Urban air mobility is an emerging transportation framework with the goal of providing on-demand, personal point-to-point transportation within and between the obstacle-rich environments of urban domains by leveraging the airspace above roads and between and above buildings and other physical infrastructure. Urban air mobility has great promise for alleviating transportation congestion, decreasing travel times, decreasing levels of pollution through the use of next-generation electric aircraft, and producing technology to benefit a wide range of industries; see Fig. 16. The purpose of this case study is to demonstrate how the component subsystems of urban air mobility have and can benefit from data science.

To achieve the promise of point-to-point, on-demand personal transportation, a number of key technological challenges must be addressed. These challenges include, but are not limited to, vehicle design, testing, and certification; vehicle fleet health monitoring; vertical garages (vertiport) design and construction for passenger on- and off-boarding; minor maintenance such as battery charging, fueling, cleaning, and major maintenance and repair; logistical scheduling systems, resembling land-based ride-shares, taxis, trains, and buses; air traffic routing, requiring next generation air traffic control; integration of autonomy for guidance, navigation, and control; and autonomous deconfliction, such as automatic dependent surveillance-broadcast (ADS-B). Nearly all of these technologies will rely on improved sensor networks, robust data communication and processing, and advanced ML algorithms. As discussed next, these system elements are at varying levels of maturity and technology readiness.

A large number of companies, both established and startup, have been exploring the development of vehicles suitable for dense and safe personal transportation in urban settings using electric propulsion. Traditional fixed wing aircraft require long runways that are spatially infeasible in cities, while helicopter designs are unable to scale to large numbers because of maneuverability and safety issues stemming from the lack of unpowered gliding capability. Innovative designs have been proposed over the past decade to achieve the needs of safe, dense operation with high maneuverability, safety, and green fuel. The designs of such vehicles require the perspective of multi-disciplinary optimization with regard to aeroservoelasticity considerations. Specifically, there is a tight physical coupling between the unsteady aerodynamics over the flexible airframe, which is further influenced by the propulsion and control surfaces. Thus, these components cannot be decoupled and designed independently, making the design space quite rich and challenging.

Historically, successful innovation of viable airframes has required years of extensive experience. The advent of modern computational capabilities will enable faster exploration of the design space through computational fluid dynamics (CFD) and finite element analysis (FEA) tools to simulate vehicle performance in a range of flight and loading conditions. These tools, and the design space over which they are leveraged, are inherently data rich. While CFD and FEA are effectively already data science tools, the potential exists for more extensive analytics to intelligently reduce the vast design space to a more manageable level. An example of how such design tools can impact outcomes can be seen in the multirotor design tool [191] that produces unexpected configurations for vehicles with multiple sets of rotating blades like a helicopter. While those designs are not likely options for passenger vehicles, the current innovation environment coupled with both the traditional design experience and modern data science tools has led to a rich array of proposed vehicles for urban operation. Options being considered include multirotors, short runway vertical takeoff and landing, and various hybrid rotor/propulsion designs. At this time, no single vehicle design is a clear frontrunner, and as with traditional fixed wing aircraft, the most likely outcome



Copyright © 2020 The Boeing Company, All Rights Reserved

Fig. 16 The urban airspace of the future will involve multiple platforms operating on various size and timescales.

will be a series of vehicle designs from companies that will share the market and the urban air space.

As with long-range air transportation, stable operation of fielded fleets of vehicles requires vehicle health monitoring, maintenance systems, and technical support. Onboard sensing and data logging capabilities for mechanical systems have been developing rapidly, with many individual subsystems (e.g., power, propulsion, cabin environment) being equipped with data systems that log and either save or regularly transmit system health information to a data server (e.g., Rolls-Royce and BMW). The sheer volume of data provided by these systems is immense, and appropriate data mining and monitoring tools are needed to flag operation and maintenance issues and to remediate issues that arise during flight. Maintenance and service for these vehicles will likely be a bit different from those for long-range flight vehicles. For aircraft powered by batteries, down time between passengers may require a quick battery charge or replacement. For vehicles with combustion engines, fuel tanks will need to be filled. Operational inspections will need to take place to ensure passenger safety. These tasks must all be performed in such a way as to facilitate efficient vehicle takeoffs, landings, and ground maneuvering. Solutions for efficiently managing these tasks are actively in development for other applications such as autonomous ocean transportation, autonomous ground transportation, and smart city infrastructure. In addition, vertical *garages* will be key to the successful adoption of urban air transportation, enabling safe takeoff and landing and facilitating efficiencies that might not be possible otherwise.

Logistical systems appropriate for handling on-demand transportation requests are quite mature. The advent of ride-share companies, such as Uber and Lyft, has driven the development of effective mobile applications that handle all aspects of the process, from user registry, payment, localization, service connection, tasking, scheduling, routing, and feedback surveys. These tools are at a level of maturity that should translate readily to the urban air mobility domain. Data science tools have been developed specifically to address these needs and are extensively leveraged for many aspects of the entire logistics system. One exception in this technology is air traffic routing. Current tools for ground transportation rely on a well-structured operational environment that has extensive physical infrastructure (e.g., roads, buildings, bridges, traffic lights, signage) and *rules-of-the-road* that inform and are informed by regulatory authorities.

A key question regarding urban air transportation is that of whether aircraft will be piloted, unpiloted, or a mix of both. Given the designs under development, the most likely scenario is the hybrid case where a mix of many different types of vehicle piloting operate simultaneously. Even in the case of human-piloted vehicles, a great deal of vehicle autonomy is in operation at any time to manage the myriad subsystems that are required for a modern flight system. The dense airspace of the future will likely exacerbate the existing challenges of autonomy. Further, the current human-centric air traffic control framework will simply be unable to address the temporal and spatial density needed for urban air mobility to be viable. However, autonomous routing, deconfliction, and obstacle avoidance are not yet at a level of maturing to be fielded. The issues are partly technological and partly policy. On the technology side, sensors and algorithms are in late stage development (e.g., ADS-B), but they have not yet been adopted into standard practice. On the policy side, work has been underway for Next Generation Air Traffic Control (NextGen ATC) for the past two decades, but final adoption has not yet reached final stages that were planned for 2025. To realize the NextGen ATC system, data science techniques are needed to provide real-time routing of all vehicles in the same way that ground-based transportation systems use.

XI. Conclusions

This paper has provided a high-level review and roadmap for the uses of data science and ML in the aerospace industry. Aerospace engineering is a data-rich endeavor, and it is poised to capitalize on the big-data and ML revolution that is redefining the modern scientific, technological, and industrial landscapes. A number of high-priority opportunities have been outlined in aerospace design,

manufacturing, testing, and services that all may be enabled by advanced data science algorithms. For example, future data-driven algorithms may enable the design of new supermaterials, streamlined flight testing for safer and more flexible designs, and enhanced manufacturing processes, including standardization, predictive assembly, and nondestructive inspection. The array of potential applications is both riveting and at times may be overwhelming. Some of these advanced technologies, such as predictive shimming and assembly, are already experiencing considerable success in commercial production. Other applications will be enabled by the digital twin of the future, with advanced optimization and learning algorithms seamlessly fusing multiphysics and multifidelity models with real-world data streams. These models will continuously improve, with potentially radical implications for the entire aerospace production pipeline. In addition to these opportunities to leverage the growing wealth of data, there are also challenges to incorporating ML into the aerospace industry. A brief summary of some high-priority opportunities and challenges is as follows:

Opportunities:

- 1) Faster design and testing cycles with digital twin and digital thread technology and accurate/efficient surrogate models
- 2) Safe, efficient, and reproducible manufacturing: predictive assembly, process control, part standardization, reproducibility, nondestructive inspection, automation, etc.
- 3) Streamlined and more reliable testing, evaluation, and certification, including anomaly detection
- 4) Improved models for complex multiscale physics, such as fluid dynamics and advanced materials and composites
- 5) Enhanced business analytics: future product development, supply chain, sales, human resources, marketing, predictive maintenance, etc.

Challenges:

- 1) Certifiable and verifiable ML models for safety-critical applications
- 2) Existence of a critical need for interpretable and generalizable ML models that incorporate partially known physics
- 3) Heterogeneity, multimodality, and multifidelity of aerospace data, which is vast in some dimensions, and sparse in others
- 4) Data mortgage, where collecting and maintaining data comes at a prohibitively high cost
- 5) Education and workforce development in classical aerospace engineering and emerging data science technologies
- 6) Fostering a culture of data sharing, open science, and reproducibility

There are clear parallels between the recent rise of data science and the transformative impact of scientific computing beginning in the 1960s. Computational science began as a specialized focus and has since evolved into a core engineering competency across many disciplines. Likewise, data science and ML proficiency will be expected as a core competency in the future workforce, highlighting the need for robust education initiatives focused on data-driven science and engineering [9].[¶] Data-driven aerospace engineering will also require changes to how teams of researchers and engineers are formed and how decisions are made. A recent *Harvard Business Review* paper [192] pointed out the need for data scientists to be integrated into the decision-making process, likening this to having Spock on the bridge in *Star Trek*. The last challenge to discuss is how to develop a *data-first* culture, which is imperative to fully harness the potential of data-enabled engineering. A data-first culture requires enterprise-wide education and adoption of principles that promote cooperation and data sharing, reproducible practices, common terminology, and an understanding of the value of data and the promise and limitations of ML algorithms. Moreover, this transformation will require ongoing curiosity and deep thinking from engineers and leaders about how processes might be improved with emerging technologies.

Current and future aircraft programs will be increasingly enabled by a wealth of data. There is low-hanging fruit in current aircraft

[¶]See <http://databookuw.com/> for video lectures, syllabi, and code that are tailored for a science and engineering audience.

programs that may provide immediate benefits with existing data. However, the full potential of data-enabled aerospace engineering will take generations of aircraft programs to realize. New programs under development, such as the new midsize airplane, will provide an opportunity to develop and test entirely new design, manufacturing, and testing capabilities. The digital twin will improve design and testing cycles through a more holistic data-driven model of critical processes and their interactions. These integrated models will rely on the integration of myriad data sources into a digital thread. Lessons learned may then be leveraged into more seamless integration of data into the aerospace industry for future programs, such as the future small aircraft. Much as information technology companies like Google are valued based on their data, aerospace giants will learn to extract value and competitive advantage by leveraging their wealth of data.

Never before has there been an architecture problem of this size, requiring the integrated efforts of information technology experts, aerospace domain engineers, and data science teams. The aerospace industry currently generates tremendous volumes of data throughout a product life cycle, but the data storage systems are not always designed to have their data extracted, much less at near real-time rates. Data must be seamless to access while maintaining security and control; software must enable analytics while performing its primary objective. Data must be integrated while being served by decades-old systems. Data exploration must be near-effortless and allow for analysis of complex systems. Analytic results must be fed back into the systems providing suggestions or predictions must be resolved without forcing all data to be migrated or new systems wholesale be replaced. There is also the risk of a data mortgage and paralysis, where more effort is spent in collecting and curating data than analyzing it. This risk motivates a shift from *big data* to *smart data*, where edge-computing and sparse/robust algorithms are used to extract key data features in real-time. Moving from a process largely grounded in storing data in local silos to openly sharing data will be difficult and will motivate entirely new incentive structures for engineers. Teams must come to identify the value they provide in part based upon the data they generate and curate.

It is also important to reiterate the need to develop ML algorithms that are specifically tailored for the aerospace industry. ML must be demystified: it is not a magic wand, but rather a growing corpus of applied optimization algorithms to build models from data. These models are only as good as the data used to train them, and great care must be taken to understand how and when these models are valid. Most ML algorithms are fundamentally interpolative, and extrapolation algorithms are both rare and challenging. Because of the need for reliable and certifiable algorithms, it is critical that physics is baked into ML algorithms. Many of the concepts related to data analytics detailed within this paper constitute novel applications that may not support a direct showing of compliance to existing regulations. Regulatory compliance is a critical challenge for emerging approaches based on data analytics. Historically, civil aviation requirements have been created and revised in response to negative events in the industry: primarily in reaction to accidents and incidents. Further, these requirements are often made under the expectation that they are deterministic in nature. As nondeterministic systems become introduced into commercial aviation systems, the basic regulatory philosophy will not support the current methods of compliance. Anywhere that data analytics are leveraged in the certification process, the same philosophical issues will need to be addressed. In place of a deterministic answer, statistical approaches, and potentially models with dis-similar architectures, or other approaches not yet identified will become a necessity. Fortunately, statistics are used in various areas of aircraft certification, related to aircraft safety assurance following single and multiple failure events.

Powered human flight is one of the greatest achievements in the history of humankind, realizing the culmination of thousands of years of science fiction, and having a profound impact on the past century. The next century of aerospace engineering will challenge us to envision and realize a new science fiction future based on the breathtaking array of new data-enhanced technologies.

Acknowledgments

We gratefully acknowledge funding support through Boeing grant 2018-ETT-PA-379. We would like to acknowledge several individuals and organizations within the University of Washington and The Boeing Company: The Boeing Advanced Research Center and its directors Per Reinhall, Santosh Devasia, and Sam Pedigo; Tia Benson Tolle, Steve Chisholm, Chris Schuller, and Todd Zarfos at Boeing; Ashis Banerjee, Bing Brunton, Ramulu Mamidala, and Navid Zobeiry at University of Washington (UW); and the eScience Institute and its directors Ed Lazowska and Magda Balazinska. We would also like to give special thanks to Mike Bragg for his support as Dean of the College of Engineering at UW and Greg Hyslop, Lynne Hopper, and Nancy Pendleton at Boeing for expanding the scope of this collaboration. We are also grateful for advice and interviews with Ryan Smith, Laura Garnett, Phil Crothers, Nia Jetter, Howard McKenzie, and Darren Macer. Finally, we are indebted to the creative efforts of Michael Erickson and the envisioning team, who translated our engineering discussions into many of the graphics in this paper.

References

- Harding, J. A., Shahbaz, M., Srinivas, M., and Kusiak, A., "Data Mining in Manufacturing: A Review," *Journal of Manufacturing Science and Engineering*, Vol. 128, No. 4, 2006, pp. 969–976. <https://doi.org/10.1115/1.2194554>
- Lynch, C., "Big Data: How Do Your Data Grow?" *Nature*, Vol. 455, No. 7209, 2008, pp. 28–29. <https://doi.org/10.1038/455028a>
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D., "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, Vol. 14, No. 1, 2008, pp. 1–37. <https://doi.org/10.1007/s10115-007-0114-2>
- Marx, V., "Biology: The Big Challenges of Big Data," *Nature*, Vol. 498, No. 7453, 2013, pp. 255–260. <https://doi.org/10.1038/498255a>
- Khoury, M. J., and Ioannidis, J. P. A., "Medicine. Big Data Meets Public Health," *Science*, Vol. 346, No. 6213, 2014, pp. 1054–1055. <https://doi.org/10.1126/science.aaa2709>
- Einav, L., and Levin, J., "Economics in the Age of Big Data," *Science*, Vol. 346, No. 6210, 2014, Paper 1243089. <https://doi.org/10.1126/science.1243089>
- Jordan, M. I., and Mitchell, T. M., "Machine Learning: Trends, Perspectives, and Prospects," *Science*, Vol. 349, No. 6245, 2015, pp. 255–260. <https://doi.org/10.1126/science.aaa8415>
- Kutz, J. N., *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*, Oxford Univ. Press, New York, 2013.
- Brunton, S. L., and Kutz, J. N., *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge Univ. Press, Cambridge, England, U.K., 2019. <https://doi.org/10.1017/9781108380690>
- Donoho, D., "50 Years of Data Science," *Journal of Computational and Graphical Statistics*, Vol. 26, No. 4, 2017, pp. 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Hey, J. G., and Trefethen, A. E., *The Data Deluge: An e-Science Perspective*, Wiley, Hoboken, NJ, 2003. <https://stewardshipgap.net/node/145>.
- Hey, A. J. G., Tansley, S., and Tolle, K. M., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA, 2009.
- Duraisamy, K., Iaccarino, G., and Xiao, H., "Turbulence Modeling in the Age of Data," *Annual Reviews of Fluid Mechanics*, Vol. 51, Jan. 2019, pp. 357–377. <https://doi.org/10.1146/annurev-fluid-010518-040547>
- Brunton, S. L., Noack, B. R., and Koumoutsakos, P., "Machine Learning for Fluid Mechanics," *Annual Review of Fluid Mechanics*, Vol. 52, Jan. 2020, pp. 477–508. <https://doi.org/10.1146/annurev-fluid-010719-060214>
- Brunton, S. L., and Kutz, J. N., "Methods for Data-Driven Multiscale Model Discovery for Materials," *Journal of Physics: Materials*, Vol. 2, No. 4, 2019, Paper 044002. <https://doi.org/10.1088/2515-7639/ab291e>

- [16] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, Cambridge, MA, 2016.
<https://doi.org/10.1007/s10710-017-9314-z>
- [17] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L., “Imagenet: A Large-Scale Hierarchical Image Database,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Inst. of Electrical and Electronics Engineers, New York, 2009, pp. 248–255.
<https://doi.org/10.1109/CVPR.2009.5206848>
- [18] Krizhevsky, A., Sutskever, I., and Hinton, G. E., “Imagenet Classification with Deep Convolutional Neural Networks,” *Communication of the Association for Computing Machinery*, Vol. 60, No. 6, May 2017, pp. 84–90.
<https://doi.org/10.1145/3065386>
- [19] Sutton, R. S., and Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2018.
- [20] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D., “A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play,” *Science*, Vol. 362, No. 6419, 2018, pp. 1140–1144.
<https://doi.org/10.1126/science.aar6404>
- [21] Hornik, K., Stinchcombe, M., and White, H., “Multilayer Feedforward Networks Are Universal Approximators,” *Neural Networks*, Vol. 2, No. 5, 1989, pp. 359–366.
[https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- [22] Loiseau, J.-C., and Brunton, S. L., “Constrained Sparse Galerkin Regression,” *Journal of Fluid Mechanics*, Vol. 838, March 2018, pp. 42–67.
<https://doi.org/10.1017/jfm.2017.823>
- [23] Raissi, M., and Karniadakis, G. E., “Hidden Physics Models: Machine Learning of Nonlinear Partial Differential Equations,” *Journal of Computational Physics*, Vol. 357, March 2018, pp. 125–141.
<https://doi.org/10.1016/j.jcp.2017.11.039>
- [24] Loiseau, J.-C., Noack, B. R., and Brunton, S. L., “Sparse Reduced-Order Modeling: Sensor-Based Dynamics to Full-State Estimation,” *Journal of Fluid Mechanics*, Vol. 844, June 2018, pp. 459–490.
<https://doi.org/10.1017/jfm.2018.147>
- [25] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., and Gulcehre, C., “Relational Inductive Biases, Deep Learning, and Graph Networks,” arXiv preprint arXiv:1806.01261, 2018.
- [26] Raissi, M., Perdikaris, P., and Karniadakis, G. E., “Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations,” *Journal of Computational Physics*, Vol. 378, Feb. 2019, pp. 686–707.
<https://doi.org/10.1016/j.jcp.2018.10.045>
- [27] Noé, F., Olsson, S., Köhler, J., and Wu, H., “Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning,” *Science*, Vol. 365, No. 6457, 2019, Paper eaaw1147.
<https://doi.org/10.1126/science.aaw1147>
- [28] Köhler, J., Klein, L., and Noé, F., “Equivariant Flows: Sampling Configurations for Multi-Body Systems with Symmetric Energies,” arXiv preprint arXiv:1910.00753, 2019.
- [29] Cranmer, M. D., Xu, R., Battaglia, P., and Ho, S., “Learning Symbolic Physics with Graph Networks,” arXiv preprint arXiv:1909.05862, 2019.
- [30] Raissi, M., Yazdani, A., and Karniadakis, G. E., “Hidden Fluid Mechanics: Learning Velocity and Pressure Fields from Flow Visualizations,” *Science*, Vol. 367, No. 6481, 2020, pp. 1026–1030.
<https://doi.org/10.1126/science.aaw4741>
- [31] Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S., “Discovering Symbolic Models from Deep Learning with Inductive Biases,” arXiv preprint arXiv:2006.11287, 2020.
- [32] Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S., “Lagrangian Neural Networks,” arXiv preprint arXiv:2003.04630, 2020.
- [33] Taira, K., Brunton, S. L., Dawson, S. T. M., Rowley, C. W., Colonius, T., McKeon, B. J., Schmidt, O. T., Gordeyev, S., Theofilis, V., and Ukeiley, L. S., “Modal Analysis of Fluid Flows: An Overview,” *AIAA Journal*, Vol. 55, No. 12, 2017, pp. 4013–4041.
<https://doi.org/10.2514/1.J056060>
- [34] Taira, K., Hemati, M. S., Brunton, S. L., Sun, Y., Duraisamy, K., Bagheri, S., Dawson, S. T. M., and Yeh, C. A., “Modal Analysis of Fluid Flows: Applications and Outlook,” *AIAA Journal*, Vol. 58, No. 3, 2020, pp. 998–1022.
<https://doi.org/10.2514/1.J058462>
- [35] Colonius, T., “An Overview of Simulation, Modeling, and Active Control of Flow/Acoustic Resonance in Open Cavities,” *39th Aerospace Sciences Meeting and Exhibit*, AIAA Paper 2001-0076, 2001.
<https://doi.org/10.2514/6.2001-76>
- [36] Kerstens, W., Pfeiffer, J., Williams, D., King, R., and Colonius, T., “Closed-Loop Control of Lift for Longitudinal Gust Suppression at Low Reynolds Numbers,” *AIAA Journal*, Vol. 49, No. 8, 2011, pp. 1721–1728.
<https://doi.org/10.2514/1.J050954>
- [37] Brunton, S. L., and Noack, B. R., “Closed-Loop Turbulence Control: Progress and Challenges,” *Applied Mechanics Reviews*, Vol. 67, No. 5, 2015, Paper 050801.
<https://doi.org/10.1115/1.4031175>
- [38] Shmilovich, A., Yadlin, Y., and Whalen, E. A., “Active Flow Control Computations: From a Single Actuator to a Complete Airplane,” *AIAA Journal*, Vol. 56, No. 12, 2018, pp. 4730–4740.
<https://doi.org/10.2514/1.J056307>
- [39] Andino, M. Y., Lin, J. C., Roman, S., Graff, E. C., Gharib, M., Whalen, E. A., and Wygnanski, I. J., “Active Flow Control on Vertical Tail Models,” *AIAA Journal*, Vol. 57, No. 8, 2019, pp. 3322–3338.
<https://doi.org/10.2514/1.J057876>
- [40] Lin, J. C., Whalen, E. A., Andino, M. Y., Graff, E. C., Lacy, D. S., Washburn, A. E., Gharib, M., and Wygnanski, I. J., “Full-Scale Testing of Active Flow Control Applied to a Vertical Tail,” *Journal of Aircraft*, Vol. 56, No. 4, 2019, pp. 1376–1386.
<https://doi.org/10.2514/1.C034907>
- [41] Hou, W., Darakananda, D., and Eldredge, J. D., “Machine Learning Based Detection of Flow Disturbances Using Surface Pressure Measurements,” *AIAA Journal*, Vol. 57, No. 12, 2019, pp. 5079–5093.
<https://doi.org/10.2514/1.J058486>
- [42] Jordan, P., and Colonius, T., “Wave Packets and Turbulent Jet Noise,” *Annual Review of Fluid Mechanics*, Vol. 45, Jan. 2013, pp. 173–195.
<https://doi.org/10.1146/annurev-fluid-011212-140756>
- [43] Spalart, P., and Allmaras, S., “A One-Equation Turbulence Model for Aerodynamic Flows,” *30th Aerospace Sciences Meeting and Exhibit*, AIAA Paper 1992-0439, 1992.
<https://doi.org/10.2514/6.1992-439>
- [44] Spalart, P. R., and Rumsey, C. L., “Effective Inflow Conditions for Turbulence Models in Aerodynamic Calculations,” *AIAA Journal*, Vol. 45, No. 10, 2007, pp. 2544–2553.
<https://doi.org/10.2514/1.29373>
- [45] Ling, J., Kurzwski, A., and Templeton, J., “Reynolds Averaged Turbulence Modelling Using Deep Neural Networks with Embedded Invariance,” *Journal of Fluid Mechanics*, Vol. 807, Nov. 2016, pp. 155–166.
<https://doi.org/10.1017/jfm.2016.615>
- [46] Singh, A. P., Medida, S., and Duraisamy, K., “Machine-Learning-Augmented Predictive Modeling of Turbulent Separated Flows over Airfoils,” *AIAA Journal*, Vol. 55, No. 7, 2017, pp. 2215–2227.
<https://doi.org/10.2514/1.J055595>
- [47] Maulik, R., San, O., Rasheed, A., and Vedula, P., “Subgrid Modelling for Two-Dimensional Turbulence Using Neural Networks,” *Journal of Fluid Mechanics*, Vol. 858, Jan. 2019, pp. 122–144.
<https://doi.org/10.1017/jfm.2018.770>
- [48] Spalart, P. R., and Garbaruk, A. V., “Correction to the Spalart–Allmaras Turbulence Model, Providing More Accurate Skin Friction,” *AIAA Journal*, Vol. 58, No. 5, 2020, pp. 1903–1905.
<https://doi.org/10.2514/1.J059489>
- [49] Thuerey, N., Weßienow, K., Prantl, L., and Hu, X., “Deep Learning Methods for Reynolds-Averaged Navier–Stokes Simulations of Airfoil Flows,” *AIAA Journal*, Vol. 58, No. 1, 2020, pp. 25–36.
<https://doi.org/10.2514/1.J058291>
- [50] Zare, A., Georgiou, T. T., and Jovanović, M. R., “Stochastic Dynamical Modeling of Turbulent Flows,” *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 3, May 2020, pp. 195–219.
<https://doi.org/10.1146/annurev-control-053018-023843>
- [51] Schmid, P. J., “Dynamic Mode Decomposition of Numerical and Experimental Data,” *Journal of Fluid Mechanics*, Vol. 656, Aug. 2010, pp. 5–0.
<https://doi.org/10.1017/S0022112010001217>
- [52] Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., and Kutz, J. N., “On Dynamic Mode Decomposition: Theory and Applications,” *Journal of Computational Dynamics*, Vol. 1, No. 2, 2014, pp. 391–421.
<https://doi.org/10.3934/jcd.2014.1.391>
- [53] Kutz, N., Brunton, S. L., Brunton, B. W., and Proctor, J. L., *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*, SIAM, Philadelphia, PA, 2016.
<https://doi.org/10.1137/1.9781611974508>

- [54] Mezić, I., “Analysis of Fluid Flows via Spectral Properties of the Koopman Operator,” *Annual Review of Fluid Mechanics*, Vol. 45, Jan. 2013, pp. 357–378.
<https://doi.org/10.1146/annurev-fluid-011212-140652>
- [55] Brunton, S. L., Brunton, B. W., Proctor, J. L., and Kutz, J. N., “Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control,” *PLoS ONE*, Vol. 11, No. 2, 2016, Paper e0150171.
<https://doi.org/10.1371/journal.pone.0150171>
- [56] McKeon, B. J., and Sharma, A. S., “A Critical Layer Model for Turbulent Pipe Flow,” *Journal of Fluid Mechanics*, Vol. 658, Sept. 2010, pp. 336–382.
<https://doi.org/10.1017/S002211201000176X>
- [57] Brunton, S. L., Proctor, J. L., and Kutz, J. N., “Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems,” *Proceedings of the National Academy of Sciences*, Vol. 113, No. 15, 2016, pp. 3932–3937.
<https://doi.org/10.1073/pnas.1517384113>
- [58] Renganathan, S. A., “Koopman-Based Approach to Nonintrusive Reduced Order Modeling: Application to Aerodynamic Shape Optimization and Uncertainty Propagation,” *AIAA Journal*, Vol. 58, No. 5, 2020, pp. 2221–2235.
<https://doi.org/10.2514/1.J058744>
- [59] Noack, B. R., Afanasiev, K., Morzynski, M., Tadmor, G., and Thiele, F., “A Hierarchy of Low-Dimensional Models for the Transient and Post-Transient Cylinder Wake,” *Journal of Fluid Mechanics*, Vol. 497, Dec. 2003, pp. 335–363.
<https://doi.org/10.1017/S0022112003006694>
- [60] Carlberg, K., Farhat, C., Cortial, J., and Amsallem, D., “The GNAT Method for Nonlinear Model Reduction: Effective Implementation and Application to Computational Fluid Dynamics and Turbulent Flows,” *Journal of Computational Physics*, Vol. 242, June 2013, pp. 623–647.
<https://doi.org/10.1016/j.jcp.2013.02.028>
- [61] Bui-Thanh, T., Willcox, K., Ghattas, O., and van Bloemen Waanders, B., “Goal-Oriented, Model-Constrained Optimization for Reduction of Large-Scale Systems,” *Journal of Computational Physics*, Vol. 224, No. 2, 2007, pp. 880–896.
<https://doi.org/10.1016/j.jcp.2006.10.026>
- [62] Bui-Thanh, T., Willcox, K., and Ghattas, O., “Model Reduction for Large-Scale Systems with High-Dimensional Parametric Input Space,” *SIAM Journal on Scientific Computing*, Vol. 30, No. 6, 2008, pp. 3270–3288.
<https://doi.org/10.1137/070694855>
- [63] Amsallem, D., Zahr, M., Choi, Y., and Farhat, C., “Design Optimization Using Hyper-Reduced-Order Models,” *Structural and Multidisciplinary Optimization*, Vol. 51, No. 4, 2015, pp. 919–940.
<https://doi.org/10.1007/s00158-014-1183-y>
- [64] Benner, P., Gugercin, S., and Willcox, K., “A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems,” *SIAM Review*, Vol. 57, No. 4, 2015, pp. 483–531.
<https://doi.org/10.1137/130932715>
- [65] Carlberg, K., Barone, M., and Antil, H., “Galerkin v. Least-Squares Petrov–Galerkin Projection in Nonlinear Model Reduction,” *Journal of Computational Physics*, Vol. 330, Feb. 2017, pp. 693–734.
<https://doi.org/10.1016/j.jcp.2016.10.033>
- [66] Carlberg, K. T., Jameson, A., Kochenderfer, M. J., Morton, J., Peng, L., and Witherden, F. D., “Recovering Missing CFD Data for High-Order Discretizations Using Deep Neural Networks and Dynamics Learning,” arXiv preprint arXiv:1812.01177, 2018.
- [67] Singh, A. P., Medida, S., and Duraisamy, K., “Machine-Learning-Augmented Predictive Modeling of Turbulent Separated Flows over Airfoils,” *AIAA Journal*, Vol. 55, No. 7, 2017, pp. 2215–2227.
<https://doi.org/10.2514/1.J055595>
- [68] Boyd, S., Boyd, S. P., and Vandenberghe, L., *Convex Optimization*, Cambridge Univ. Press, Cambridge, England, U.K., 2004.
- [69] Kalman, E., “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Fluids Engineering*, Vol. 82, No. 1, 1960, pp. 35–45.
<https://doi.org/10.1115/1.3662552>
- [70] Bertsekas, D. P., *Dynamic Programming and Optimal Control*, Vol. 1, Athena Scientific, Belmont, MA, 1995.
- [71] Rao, C. V., Wright, S. J., and Rawlings, J. B., “Application of Interior-Point Methods to Model Predictive Control,” *Journal of Optimization Theory and Applications*, Vol. 99, No. 3, 1998, pp. 723–757.
<https://doi.org/10.1023/A:1021711402723>
- [72] Dullerud, G. E., and Fernando Paganini, F., “A Course in Robust Control Theory: A Convex Approach,” *Texts in Applied Mathematics*, Springer, Berlin, 2000.
- [73] Skogestad, S., and Postlethwaite, I., *Multivariable Feedback Control: Analysis and Design*, 2nd ed., Wiley, Hoboken, NJ, 2005.
- [74] Joshi, S., and Boyd, S., “Sensor Selection via Convex Optimization,” *IEEE Transactions on Signal Processing*, Vol. 57, No. 2, 2009, pp. 451–462.
<https://doi.org/10.1109/TSP.2008.2007095>
- [75] Aravkin, A., Burke, J. V., Ljung, L., Lozano, A., and Pillonetto, G., “Generalized Kalman Smoothing: Modeling and Algorithms,” *Automatica*, Vol. 86, Dec. 2017, pp. 63–86.
<https://doi.org/10.1016/j.automatica.2017.08.011>
- [76] Manohar, K., Brunton, B. W., Kutz, J. N., and Brunton, S. L., “Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns,” *IEEE Control Systems Magazine*, Vol. 38, No. 3, 2018, pp. 63–86.
<https://doi.org/10.1109/MCS.2018.2810460>
- [77] Bottou, L., Curtis, F. E., and Nocedal, J., “Optimization Methods for Large-Scale Machine Learning,” *SIAM Review*, Vol. 60, No. 2, 2018, pp. 223–311.
<https://doi.org/10.1137/16M1080173>
- [78] Shapiro, A., Dentcheva, D., and Ruszczyński, A., *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, PA, 2009.
<https://doi.org/10.1137/1.9781611973433>
- [79] Tsitsiklis, J., Bertsekas, D., and Athans, M., “Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms,” *IEEE Transactions on Automatic Control*, Vol. 31, No. 9, 1986, pp. 803–812.
<https://doi.org/10.1109/TAC.1986.1104412>
- [80] Bottou, L., “Large-Scale Machine Learning with Stochastic Gradient Descent,” *Proceedings of COMPSTAT 2010*, edited by Y. Lechevallier, and G. Saporta, Springer-Verlag, Berlin, 2010, pp. 177–186.
https://doi.org/10.1007/978-3-7908-2604-3_16
- [81] Kingma, D. P., and Ba, J., “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.
- [82] Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D., “Stochastic Subgradient Method Converges on Tame Functions,” *Foundations of Computational Mathematics*, Vol. 31, No. 9, 2018, pp. 1–36.
<https://doi.org/10.1007/s10208-018-09409-5>
- [83] Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J., “Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization,” *Advances in Neural Information Processing Systems*, Vol. 29, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Curran Associates, Inc., 2016, pp. 1145–1153.
<https://doi.org/10.5555/3157096.3157225>
- [84] Aravkin, A., and Davis, D., “Trimmed Statistical Estimation via Variance Reduction,” *Mathematics of Operations Research*, Vol. 45, No. 1, 2020, pp. 292–322.
<https://doi.org/10.1287/moor.2019.0992>
- [85] Rockafellar, R. T., and Wets, R. J.-B., *Variational Analysis*, Vol. 317, Springer Science & Business Media, Berlin, 2009.
- [86] Burke, J. V., and Ferris, M. C., “A Gauss–Newton Method for Convex Composite Optimization,” *Mathematical Programming*, Vol. 71, No. 2, 1995, pp. 179–194.
<https://doi.org/10.1007/BF01585997>
- [87] Davis, D., and Drusvyatskiy, D., “Stochastic Model-Based Minimization of Weakly Convex Functions,” *SIAM Journal on Optimization*, Vol. 29, No. 1, 2019, pp. 207–239.
<https://doi.org/10.1137/18M1178244>
- [88] Golub, G., and Pereyra, V., “Separable Nonlinear Least Squares: The Variable Projection Method and Its Applications,” *Inverse problems*, Vol. 19, No. 2, 2003, Paper R1.
<https://doi.org/10.1088/0266-5611/19/2/201>
- [89] Zheng, P., Askham, T., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y., “A Unified Framework for Sparse Related Regularized Regression: SR3,” *IEEE Access*, Vol. 7, No. 1, 2019, pp. 1404–1423.
<https://doi.org/10.1109/ACCESS.2018.2886528>
- [90] Nocedal, J., and Wright, S., *Numerical Optimization*, Springer-Verlag, New York, 2006.
- [91] Combettes, P. L., and Pesquet, J.-C., “Proximal Splitting Methods in Signal Processing,” *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, New York, 2011, pp. 185–212.
https://doi.org/10.1007/978-1-4419-9569-8_10
- [92] Parikh, N., and Boyd, S., “Proximal Algorithms,” *Foundations and Trends® in Optimization*, Vol. 1, No. 3, 2014, pp. 127–239.
<https://doi.org/10.1561/24000000003>
- [93] Attouch, H., Bolte, J., and Svaiter, B. F., “Convergence of Descent Methods for Semi-Algebraic and Tame Problems: Proximal Algorithms, Forward–Backward Splitting, and Regularized Gauss–Seidel Methods,” *Mathematical Programming*, Vol. 137, Nos. 1–2, 2013,

- pp. 91–129.
<https://doi.org/10.1007/s10107-011-0484-9>
- [94] Michael, W., “Mahoney and Petros Drineas CUR Matrix Decompositions for Improved Data Analysis,” *Proceedings of the National Academy of Sciences*, Vol. 106, No. 3, 2009, pp. 697–702.
<https://doi.org/10.1073/pnas.0803205106>
- [95] Mahoney, M. W., “Randomized Algorithms for Matrices and Data,” *Foundations and Trends in Machine Learning*, Vol. 3, No. 3, 2011, pp. 123–224.
<https://doi.org/10.1561/22000000035>
- [96] Halko, N., Martinsson, P. G., and Tropp, J. A., “Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions,” *SIAM Review*, Vol. 53, No. 2, 2011, pp. 217–288.
<https://doi.org/10.1137/090771806>
- [97] Liberty, E., “Simple and Deterministic Matrix Sketching,” *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD’13, Assoc. for Computing Machinery, New York, 2013, pp. 581–588.
<https://doi.org/10.1145/2487575.2487623>
- [98] Erichson, N. B., Voronin, S., Brunton, S. L., and Kutz, J. N., “Randomized Matrix Decompositions Using R,” *Journal of Statistical Software*, Vol. 89, No. 1, 2019, pp. 1–48.
<https://doi.org/10.18637/jss.v089.i11>
- [99] Sarlos, T., “Improved Approximation Algorithms for Large Matrices via Random Projections,” *47th Annual IEEE Symposium on Foundations of Computer Science*, 2006, pp. 143–152.
<https://doi.org/10.1109/FOCS.2006.37>
- [100] Martinsson, P. G., Rokhlin, V., and Tytgert, M., “A Randomized Algorithm for the Decomposition of Matrices,” *Applied and Computational Harmonic Analysis*, Vol. 30, No. 1, 2011, pp. 47–68.
<https://doi.org/10.1016/j.acha.2010.02.003>
- [101] Rokhlin, V., Szlam, A., and Tytgert, M., “A Randomized Algorithm for Principal Component Analysis,” *SIAM Journal on Matrix Analysis and Applications*, Vol. 31 No. 3, 2010, pp. 1100–1124.
<https://doi.org/10.1137/080736417>
- [102] Halko, N., Martinsson, P. G., Shkolnisky, Y., and Tytgert, M., “An Algorithm for the Principal Component Analysis of Large Data Sets,” *SIAM Journal on Scientific Computing*, Vol. 33, No. 5, 2011, pp. 2580–2594.
<https://doi.org/10.1137/100804139>
- [103] Duersch, J. A., and Gu, M., “Randomized QR with Column Pivoting,” *SIAM Journal on Scientific Computing*, Vol. 39, No. 4, 2017, pp. C263–C291.
<https://doi.org/10.1137/15M1044680>
- [104] Shabat, G., Shmueli, Y., Aizenbud, Y., and Averbuch, A., “Randomized LU Decomposition,” *Applied and Computational Harmonic Analysis*, Vol. 44, No. 2, 2018, pp. 246–272.
<https://doi.org/10.1016/j.acha.2016.04.006>
- [105] Erichson, N. B., Manohar, K., Brunton, S. L., and Kutz, J. N., “Randomized CP Tensor Decomposition,” *Machine Learning: Science and Technology*, Vol. 1, No. 2, 2020, Paper 025012.
<https://doi.org/10.1088/2632-2153/ab8240>
- [106] Erichson, N. B., Mathelin, L., Kutz, J. N., and Brunton, S. L., “Randomized Dynamic Mode Decomposition,” *SIAM Journal on Applied Dynamical Systems*, Vol. 18, No. 4, 2019, pp. 1867–1891.
<https://doi.org/10.1137/18M1215013>
- [107] Alla, A., and Kutz, J. N., “Randomized Model Order Reduction,” *Advances in Computational Mathematics*, Vol. 45, No. 3, 2019, pp. 1251–1271.
<https://doi.org/10.1007/s10444-018-09655-9>
- [108] Bai, Z., Kaiser, E., Proctor, J. L., Kutz, J. N., and Brunton, S. L., “Dynamic Mode Decomposition for Compressive System Identification,” *AIAA Journal*, Vol. 58, No. 2, 2020, pp. 561–574.
<https://doi.org/10.2514/1.J057870>
- [109] Huber, P. J., “John W. Tukey’s Contributions to Robust Statistics,” *Annals of Statistics*, Vol. 30, No. 6, 2002, pp. 1640–1648.
<https://doi.org/10.1214/aos/1043351251>
- [110] Candès, E. J., Li, X., Ma, Y., and Wright, J., “Robust Principal Component Analysis?” *Journal of the ACM*, Vol. 58, No. 3, 2011, Paper 11.
<https://doi.org/10.1145/1970392.1970395>
- [111] Gavish, M., and Donoho, D. L., “The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$,” *IEEE Transactions on Information Theory*, Vol. 60, No. 8, 2014, pp. 5040–5053.
<https://doi.org/10.1109/TIT.2014.2323359>
- [112] Bouwmans, T., and Zahzah, E. H., “Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance,” *Computer Vision and Image Understanding*, Vol. 122, May 2014, pp. 22–34.
<https://doi.org/10.1016/j.cviu.2013.11.009>
- [113] Kondor, D., Csabai, I., Dobos, L., Szüle, J., Barankai, N., Hanyecz, T., Sebök, T., Kallus, Z., and Vattay, G., “Using Robust PCA to Estimate Regional Characteristics of Language use from Geo-Tagged Twitter Messages,” *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, Inst. of Electrical and Electronics Engineers, New York, 2013, pp. 393–398.
<https://doi.org/10.1109/CogInfoCom.2013.6719277>
- [114] Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y., “Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization,” *Advances in Neural Information Processing Systems*, Vol. 22, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Curran Associates, Inc., 2009, pp. 2080–2088, <https://proceedings.neurips.cc/paper/2009/file/c45147dee729311ef5b5c3003946c48f-Paper.pdf>.
- [115] Boschert, S., and Rosen, R., “Digital Twin—The Simulation Aspect,” *Mechatronic Futures*, Springer, Cham, Switzerland, 2016, pp. 59–74.
https://doi.org/10.1007/978-3-319-32156-1_5
- [116] Grieves, M., and Vickers, J., “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems,” *Transdisciplinary Perspectives on Complex Systems*, Springer, Cham, Switzerland, 2017, pp. 85–113.
https://doi.org/10.1007/978-3-319-38756-7_4
- [117] Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., and Sui, F., “Digital Twin-Driven Product Design, Manufacturing and Service with Big Data,” *International Journal of Advanced Manufacturing Technology*, Vol. 94, Nos. 9–12, 2018, pp. 3563–3576.
<https://doi.org/10.1007/s00170-017-0233-1>
- [118] Rasheed, A., San, O., and Kvamsdal, T., “Digital Twin: Values, Challenges and Enablers,” arXiv preprint arXiv:1910.01719, 2019.
- [119] Chinesta, F., Cueto, E., Abisset-Chavanne, E., Duval, J. L., and Khaldi, F. E., “Virtual, Digital and Hybrid Twins: A New Paradigm in Data-Based Engineering and Engineered Data,” *Archives of Computational Methods in Engineering*, Vol. 27, No. 1, 2020, pp. 105–134.
<https://doi.org/10.1007/s11831-018-9301-4>
- [120] Everson, R., and Sirovich, L., “Karhunen–Loève Procedure for Gappy Data,” *Journal of the Optical Society of America A*, Vol. 12, No. 8, 1995, pp. 1657–1664.
<https://doi.org/10.1364/JOSAA.12.001657>
- [121] Willcox, K., “Unsteady Flow Sensing and Estimation via the Gappy Proper Orthogonal Decomposition,” *Computers & Fluids*, Vol. 35, No. 2, 2006, pp. 208–226.
<https://doi.org/10.1016/j.compfluid.2004.11.006>
- [122] Maxime Barrault, M., Maday, Y., Nguyen, N. C., and Patera, A. T., “An ‘Empirical Interpolation’ Method: Application to Efficient Reduced-Basis Discretization of Partial Differential Equations,” *Comptes Rendus Mathématique*, Vol. 339, No. 9, 2004, pp. 667–672.
<https://doi.org/10.1016/j.crma.2004.08.006>
- [123] Chaturantabut, S., and Sorensen, D. C., “Nonlinear Model Reduction via Discrete Empirical Interpolation,” *SIAM Journal on Scientific Computing*, Vol. 32, No. 5, 2010, pp. 2737–2764.
<https://doi.org/10.1137/090766498>
- [124] Drmac, Z., and Gugercin, S., “A New Selection Operator for the Discrete Empirical Interpolation Method—Improved a Priori Error Bound and Extensions,” *SIAM Journal on Scientific Computing*, Vol. 38, No. 2, 2016, pp. A631–A648.
<https://doi.org/10.1137/15M1019271>
- [125] Candès, E. J., “Compressive Sampling,” *Proceedings of the International Congress of Mathematicians*, Vol. 3, Aug. 2006, pp. 1433–1452.
<https://doi.org/10.4171/022-3/69>
- [126] Donoho, D. L., “Compressed Sensing,” *IEEE Transactions on Information Theory*, Vol. 52, No. 4, 2006, pp. 1289–1306.
<https://doi.org/10.1109/TIT.2006.871582>
- [127] Candès, E. J., Romberg, J., and Tao, T., “Stable Signal Recovery from Incomplete and Inaccurate Measurements,” *Communications in Pure and Applied Mathematics*, Vol. 59, No. 8, 2006, pp. 1207–1223.
<https://doi.org/10.1002/cpa.20124>
- [128] Candès, E. J., and Tao, T., “Near Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?” *IEEE Transactions on Information Theory*, Vol. 52, No. 12, 2006, pp. 5406–5425.
<https://doi.org/10.1109/TIT.2006.885507>
- [129] Candès, E. J., Romberg, J., and Tao, T., “Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information,” *IEEE Transactions on Information Theory*, Vol. 52, No. 2, 2006, pp. 489–509.
<https://doi.org/10.1109/TIT.2005.862083>

- [130] Baraniuk, R. G., "Compressive Sensing," *IEEE Signal Processing Magazine*, Vol. 24, No. 4, 2007, pp. 118–121. <https://doi.org/10.1109/MSP.2007.4286571>
- [131] Dhingra, N. K., Jovanovic, M. R., and Luo, Z.-Q., "An ADMM Algorithm for Optimal Sensor and Actuator Selection," *53rd IEEE Conference on Decision and Control*, 2014, pp. 4039–4044. <https://doi.org/10.1109/CDC.2014.7040017>
- [132] Manohar, K., Brunton, S. L., and Kutz, J. N., "Environmental Identification in Flight Using Sparse Approximation of Wing Strain," *Journal of Fluids and Structures*, Vol. 70, April 2017, pp. 162–180. <https://doi.org/10.1016/j.jfluidstructs.2017.01.008>
- [133] Maute, K., Nikbay, M., and Farhat, C., "Coupled Analytical Sensitivity Analysis and Optimization of Three-Dimensional Nonlinear Aeroelastic Systems," *AIAA Journal*, Vol. 39, No. 11, 2001, pp. 2051–2061. <https://doi.org/10.2514/2.1227>
- [134] Dupuis, R., Jouhaud, J.-C., and Sagaut, P., "Surrogate Modeling of Aerodynamic Simulations for Multiple Operating Conditions Using Machine Learning," *AIAA Journal*, Vol. 56, No. 9, 2018, pp. 3622–3635. <https://doi.org/10.2514/1.J056405>
- [135] Swischuk, R., Kramer, B., Huang, C., and Willcox, K., "Learning Physics-Based Reduced-Order Models for a Single-Injector Combustion Process," *AIAA Journal*, Vol. 58, No. 6, 2020, pp. 2658–2672. <https://doi.org/10.2514/1.J058943>
- [136] Bongard, J., and Lipson, H., "Automated Reverse Engineering of Nonlinear Dynamical Systems," *Proceedings of the National Academy of Sciences*, Vol. 104, No. 24, 2007, pp. 9943–9948. <https://doi.org/10.1073/pnas.0609476104>
- [137] Schmidt, M., and Lipson, H., "Distilling Free-Form Natural Laws from Experimental Data," *Science*, Vol. 324, No. 5923, 2009, pp. 81–85. <https://doi.org/10.1126/science.1165893>
- [138] Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E., and Kutz, J. N., "Chaos as an Intermittently Forced Linear System," *Nature Communications*, Vol. 8, No. 19, 2017, pp. 1–9. <https://doi.org/10.1038/s41467-017-00030-8>
- [139] Lee, K., and Carlberg, K. T., "Model Reduction of Dynamical Systems on Nonlinear Manifolds Using Deep Convolutional Autoencoders," *Journal of Computational Physics*, Vol. 404, March 2020, Paper 108973. <https://doi.org/10.1016/j.jcp.2019.108973>
- [140] Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L., "Data-Driven Discovery of Coordinates and Governing Equations," *Proceedings of the National Academy of Sciences*, Vol. 116, No. 45, 2019, pp. 22,445–22,451. <https://doi.org/10.1073/pnas.1906995116>
- [141] Kennedy, M. C., and O'Hagan, A., "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 3, Aug. 2001, pp. 425–464. <https://doi.org/10.1111/1467-9868.00294>
- [142] Arendt, P., Apley, D. W., and Chen, W., "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability," *Journal of Mechanical Design*, Vol. 134, No. 10, 2012, Paper 100908. <https://doi.org/10.1115/1.4007390>
- [143] Quiñero-Candela, J., and Rasmussen, C. E., "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, Vol. 6, Dec. 2005, pp. 1939–1959. <https://doi.org/10.5555/1046920.1194909>
- [144] Kaheman, K., Kaiser, E., Strom, B., Kutz, J. N., and Brunton, S. L., "Learning Discrepancy Models from Experimental Data," arXiv preprint arXiv:1909.08574, 2019.
- [145] Raymer, D., *Aircraft Design: A Conceptual Approach*, AIAA, Reston, VA, 2012. <https://doi.org/10.2514/4.104909>
- [146] Oswald, W. B., "The Transverse Force Distribution on Ellipsoidal and Nearly Ellipsoidal Bodies Moving in an Arbitrary Potential Flow," Ph.D. Thesis, California Inst. of Technology, Pasadena, CA, 1932. <https://doi.org/10.7907/QK4H-C181>
- [147] Oswald, W. B., "General Formulas and Charts for the Calculation of Airplane Performance," NACA TR-408, 1932.
- [148] Cramer, E. J., Dennis, J. E., Jr., Frank, P. D., Lewis, R. M., and Shubin, G. R., "Problem Formulation for Multidisciplinary Optimization," *SIAM Journal on Optimization*, Vol. 4, No. 4, 1994, pp. 754–776. <https://doi.org/10.1137/0804044>
- [149] Booker, A. J., Dennis, J. E., Frank, P. D., Serafini, D. B., Torczon, V., and Trosset, M. W., "A Rigorous Framework for Optimization of Expensive Functions by Surrogates," *Structural Optimization*, Vol. 17, No. 1, 1999, pp. 1–13. <https://doi.org/10.1007/BF01197708>
- [150] Mavris, D. N., Bandte, O., and DeLaurentis, D. A., "Robust Design Simulation: A Probabilistic Approach to Multidisciplinary Design," *Journal of Aircraft*, Vol. 36, No. 1, 1999, pp. 298–307. <https://doi.org/10.2514/2.2437>
- [151] Bowcutt, K., Kuruwila, G., Grandine, T. A., Hogan, T. A., and Cramer, E. J., "Advancements in Multidisciplinary Design Optimization Applied to Hypersonic Vehicles to Achieve Closure," *15th International Space Planes and Hypersonic Systems and Technologies Conference*, AIAA Paper 2008-2591, 2008. <https://doi.org/10.2514/6.2008-2591>
- [152] Rallabhandi, S. K., and Mavris, D. N., "Simultaneous Airframe and Propulsion Cycle Optimization for Supersonic Aircraft Design," *Journal of Aircraft*, Vol. 45, No. 1, 2008, pp. 38–55. <https://doi.org/10.2514/1.33183>
- [153] Manan, A., and Cooper, J., "Design of Composite Wings Including Uncertainties: A Probabilistic Approach," *Journal of Aircraft*, Vol. 46, No. 2, 2009, pp. 601–607. <https://doi.org/10.2514/1.39138>
- [154] Henderson, R. P., Martins, J. R. A. A., and Perez, R. E., "Aircraft Conceptual Design for Optimal Environmental Performance," *Aeronautical Journal*, Vol. 116, No. 1175, 2012, pp. 1–22. <https://doi.org/10.1017/S000192400000659X>
- [155] Martins, J. R. A. A., and Lambe, A., "Multidisciplinary Design Optimization: A Survey of Architectures," *AIAA Journal*, Vol. 51, No. 9, 2013, pp. 2049–2075. <https://doi.org/10.2514/1.J051895>
- [156] Bons, N., and Martins, J. R. A. A., "Aerostructural Wing Design Exploration with Multidisciplinary Design Optimization," *AIAA Scitech 2020 Forum*, AIAA Paper 2020-0544, 2020. <https://doi.org/10.2514/6.2020-0544>
- [157] Baran, I., Cinar, K., Ersoy, N., Akkerman, R., and Hattel, J. H., "A Review on the Mechanical Modeling of Composite Manufacturing Processes," *Archives of Computational Methods in Engineering*, Vol. 24, No. 2, 2017, pp. 365–395. <https://doi.org/10.1007/s11831-016-9167-2>
- [158] Singh, S., Shebab, E., Higgins, N., Fowler, K., Tomiyama, T., and Fowler, C., "Challenges of Digital Twin in High Value Manufacturing," SAE TP 2018-01-1928, 2018. <https://doi.org/10.4271/2018-01-1928>
- [159] Manohar, K., Hogan, T., Buttrick, J., Banerjee, A. G., Kutz, J. N., and Brunton, S. L., "Predicting Shim Gaps in Aircraft Assembly with Machine Learning and Sparse Sensing," *Journal of Manufacturing Systems*, Vol. 48, July 2018, pp. 87–95. <https://doi.org/10.1016/j.jmsy.2018.01.011>
- [160] Conduit, B. D., Jones, N. G., Stone, H. J., and Conduit, G. J., "Design of a Nickel-Base Superalloy Using a Neural Network," *Materials & Design*, Vol. 131, Oct. 2017, pp. 358–365. <https://doi.org/10.1016/j.matdes.2017.06.007>
- [161] Verpoort, P. C., MacDonald, P., and Conduit, G. J., "Materials Data Validation and Imputation with an Artificial Neural Network," *Computational Materials Science*, Vol. 147, May 2018, pp. 176–185. <https://doi.org/10.1016/j.commatsci.2018.02.002>
- [162] Conduit, B. D., Jones, N. G., Stone, H. J., and Conduit, G. J., "Probabilistic Design of a Molybdenum-Base Alloy Using a Neural Network," *Scripta Materialia*, Vol. 146, March 2018, pp. 82–86. <https://doi.org/10.1016/j.scriptamat.2017.11.008>
- [163] Green, A. G., Conduit, G., and Krüger, F., "Quantum Order-by-Disorder in Strongly Correlated Metals," *Annual Review of Condensed Matter Physics*, Vol. 9, March 2018, pp. 59–77. <https://doi.org/10.1146/annurev-conmatphys-033117-053925>
- [164] Sachs, U., "Friction and Bending in Thermoplastic Composites Forming Processes," Ph.D. Thesis, Univ. of Twente, Enschede, The Netherlands, 2014.
- [165] Hannappel, S. P., "Forming of UD Fibre Reinforced Thermoplastics," Ph.D. Thesis, Univ. of Twente, Enschede, The Netherlands, 2013.
- [166] Juarez, P., Gregory, E., and Cramer, K., "In Situ Thermal Inspection of Automated Fiber Placement Manufacturing," *SAMPE Journal*, Vol. 2102, No. 1, 2020, Paper 120005. <https://doi.org/10.1063/1.5099847>
- [167] Cemenska, J., Rudberg, T., and Henscheid, M., "Automated In-Process Inspection System for AFP Machines," *SAE International Journal of Aerospace*, Vol. 8, Sept. 2015, pp. 303–309. <https://doi.org/10.4271/2015-01-2608>
- [168] Sacco, C., Radwan, A. B., Beatty, T., and Harik, R., "Machine Learning Based AFP Inspection: A Tool for Characterization and Integration," *SAMPE Journal*, 2020. <https://doi.org/10.33599/nasampe/s.19.1594>
- [169] Blom, A. W., "Structural Performance of Fiber-Placed, Variable-Stiffness Composite Conical and Cylindrical Shells," Ph.D. Thesis, Delft Univ. of Technology, Delft, The Netherlands, 2010.

- [170] Cohn, D. A., Ghahramani, Z., and Jordan, M. I., "Active Learning with Statistical Models," *Journal of Artificial Intelligence Research*, Vol. 4, March 1996, pp. 129–145.
<https://doi.org/10.1613/jair.295>
- [171] Settles, B., "Active Learning Literature Survey," Dept. of Computer Sciences, Univ. of Wisconsin-Madison TR 1648, 2009.
- [172] de Silva, B. M., Callahan, J., Jonker, J., Goebel, N., Klemisch, J., McDonald, D., Hicks, N., Kutz, J. N., Brunton, S. L., and Aravkin, A. Y., "Physics-Informed Machine Learning for Sensor Fault Detection with Flight Test Data," arXiv preprint arXiv:2006.13380, 2020.
- [173] Williams, M. O., Rowley, C. W., Mezić, I., and Kevrekidis, I. G., "Data Fusion via Intrinsic Dynamic Variables: An Application of Data-Driven Koopman Spectral Analysis," *EPL*, Vol. 109, No. 4, 2015, Paper 40007.
<https://doi.org/10.1209/0295-5075/109/40007>
- [174] Marsh, B. J., "Laser Tracker Assisted Aircraft Machining and Assembly," SAE TP 2008-01-2313, 2008.
<https://doi.org/10.4271/2008-01-2313>
- [175] Jamshidi, J., Kayani, A., Iravani, P., Maropoulos, P. G., and Summers, M. D., "Manufacturing and Assembly Automation by Integrated Metrology Systems for Aircraft Wing Fabrication," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, Vol. 224, No. 1, 2010, pp. 25–36.
<https://doi.org/10.1243/09544054JEM1280>
- [176] Muelaner, J. E., and Maropoulos, P. G., "Design for Measurement Assisted Determinate Assembly (MADA) of Large Composite Structures," *Journal of the Coordinate Metrology Systems Conference*, 2010.
- [177] Marsh, B. J., Vanderwiell, T., VanScotter, K., and Thompson, M., "Method for Fitting Part Assemblies," U.S. Patent US7756321B2, 2010.
- [178] Muelaner, J. E., Kayani, A., Martin, O., and Maropoulos, P., "Measurement Assisted Assembly and the Roadmap to Part-to-Part Assembly," *Proceedings of DET2011 7th International Conference on Digital Enterprise Technology*, Univ. of Bath, Sept. 2011, pp. 11–19.
- [179] Chouvion, B., Popov, A., Ratchev, S., Mason, C., and Summers, M., "Interface Management in Wing-Box Assembly," SAE TP 2011-01-2640, 2011.
<https://doi.org/10.4271/2011-01-2640>
- [180] Muelaner, J. E., and Maropoulos, P., "Integrated Dimensional Variation Management in the Digital Factory," *Proceedings of DET2011 7th International Conference on Digital Enterprise Technology*, Univ. of Bath, Sept. 2011, pp. 39–46.
- [181] Muelaner, J. E., Martin, O. C., and Maropoulos, P. G., "Achieving Low Cost and High Quality Aero Structure Assembly Through Integrated Digital Metrology Systems," *Procedia CIRP*, Vol. 7, Jan. 2013, pp. 688–693.
<https://doi.org/10.1016/j.procir.2013.06.054>
- [182] Boyd-Davis, T. M., Jones, D. D., and Zimmerman, T. E., "Digitally Designed Shims for Joining Parts of an Assembly," U.S. Patent US9429935B2, 2014.
- [183] Vasquez, C. M., Boyd-Davis, T. M., Valenzuela, D. I., and Jones, D. D., "Systems and Methods for Robotic Measurement of Parts," U.S. Patent US9958854B2, 2014.
- [184] Valenzuela, D. I., Boyd-Davis, T. M., and Jones, D. D., "Systems, Methods, and Apparatus for Automated Predictive Shimming for Large Structures," U.S. Patent US9599983B2, 2015.
- [185] Boyd-Davis, T. M., Jones, D. D., and Zimmerman, T. E., "Methods of Fabricating Shims for Joining Parts," U.S. Patent US9429935B2, Aug. 30, 2016.
- [186] Antolin-Urbaneja, J. C., Livinalli, J., Puerto, M., Liceaga, M., Rubio, A., San-Roman, A., and Goenaga, I., "End-Effector for Automatic Shimming of Composites," SAE TP 2016-01-2111, 2016.
<https://doi.org/10.4271/2016-01-2111>
- [187] Golub, G., and Kahan, W., "Calculating the Singular Values and Pseudo-Inverse of a Matrix," *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, Vol. 2, No. 2, 1965, pp. 205–224.
<https://doi.org/10.1137/0702016>
- [188] Brunton, B. W., Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Sparse Sensor Placement Optimization for Classification," *SIAM Journal on Applied Mathematics*, Vol. 76, No. 5, 2016, pp. 2099–2122.
<https://doi.org/10.1137/15M1036713>
- [189] Johnson, W., "Model for Vortex Ring State Influence on Rotorcraft Flight Dynamics," NASA TP-2005-213477, 2005.
- [190] Gray, G. J., "Report of the Panel to Review the V-22 Program," Tech. Rept., Dept. of Defense Panel Review, 2001.
- [191] Du, T., Schulz, A., Zhu, B., Bickel, B., and Matusik, W., "Computational Multicopter Design," *ACM Transactions*, Vol. 35, No. 6, 2016.
<https://doi.org/10.1145/2980179.2982427>
- [192] Patil, T. H. D. J., and Davenport, T., "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, Vol. 90, No. 10, 2012, pp. 70–76.

P. Givi
Associate Editor