Mechanistic Data Science for Engineering

Draft: An Introduction

By

Wing Kam Liu, w-liu@northwestern.edu

Walter P. Murphy Professor and Director of Global Center on Advanced Material Systems and Simulation

Zhengtao Gan, zhengtao.gan@northwestern.edu

Research Assistant Professor

Mark Fleming, mark.fleming@northwestern.edu

Adjunct Professor, Northwestern University, Principal Engineer, Fusion Engineering, LLC

Northwestern University, Department of Mechanical Engineering

Table of Contents

Preface

This book presents Mechanistic Data Science (MDS) as a structured methodology for coupling data with mathematics and scientific principles to solve intractable problems. Dictionary.com defines the word "mechanistic" as referring to theories which explain a phenomenon in purely physical or deterministic terms. Traditional data science methodologies require copious quantities of data to show a reliable pattern which can be used to draw conclusions. The amount of data required to find a solution can be greatly reduced by considering the mathematical science principles. Although in reality there is a broad spectrum for the mix of data and fundamentals, for this book, we focus on three types. Type 1 is a problem with abundant data but undeveloped or unavailable fundamental principles, often called a *purely data-driven* problem. Type 2 is a problem with *limited data and scientific knowledge*, and neither the data nor the scientific principles provide a complete solution. Type 3, *known mathematical science principles with uncertain parameters*, which can be computationally burdensome to solve.

Mechanistic data science is an innovative methodology which can be the key to addressing problems that previously could only be dealt with through trial and error or experience. It works by combining available data with the fundamental principles of mathematical science through neural networks. The first MDS challenge addressed is the generation and collection of multimodal data (such as testing, simulations, or databases). Using the data collected, we will show how to extract mechanistic features using basic mathematical tools, including continuous and discrete/digital analysis. Next, we will show how to perform knowledge-driven dimension reduction to streamline the analysis. Reduced order surrogate models will be created to introduce the fundamental physics into the solution of the problem. Basic mathematical tools will be used for this, including Fourier analysis, regression, continuous and discrete mathematics, and image analysis. Neural networks for regression and classification will be performed. These data and mechanistic analysis steps will be coupled for the system and design.

This book is written in a spectral style that is useful to high school students and teachers, engineering and data science undergraduate and graduate students, as well as practicing scientists and engineers. The book was initiated based on the notes of a Northwestern University summer course taught to high school students and engineering undergraduate students. Numerous examples are given to describe in-depth fundamental concepts in terms of everyday terminology. Our end goal of this book is to provide our readers with a mechanistic data science methodology to solve problems and make decisions by combining available data and mathematical science, not just to write another data science textbook.

1. Introduction

Most challenging problems require a combination of data and scientific or underlying principles to find a solution. The power of mechanistic data science is that the techniques which will be shown in this book apply to problems in the physical sciences and mathematics, as well as manufacturing, medical, social science, and business. As shown in Figure 1, mechanistic data science combines established equations from mathematical science with data and measurements through techniques such as neural networks to address problems which were previously intractable. Although the mechanistic data science methodology will be mostly described in terms of engineering and science examples, the same methodology is applicable to any other walk of life in which data is available and decisions are required.

This book is presented in a manner which will be applicable to high school level students and teachers, college students and professors, and working professionals. Sections that are more advanced will be noted to notify the reader that additional background knowledge may be required to fully understand that particular section.

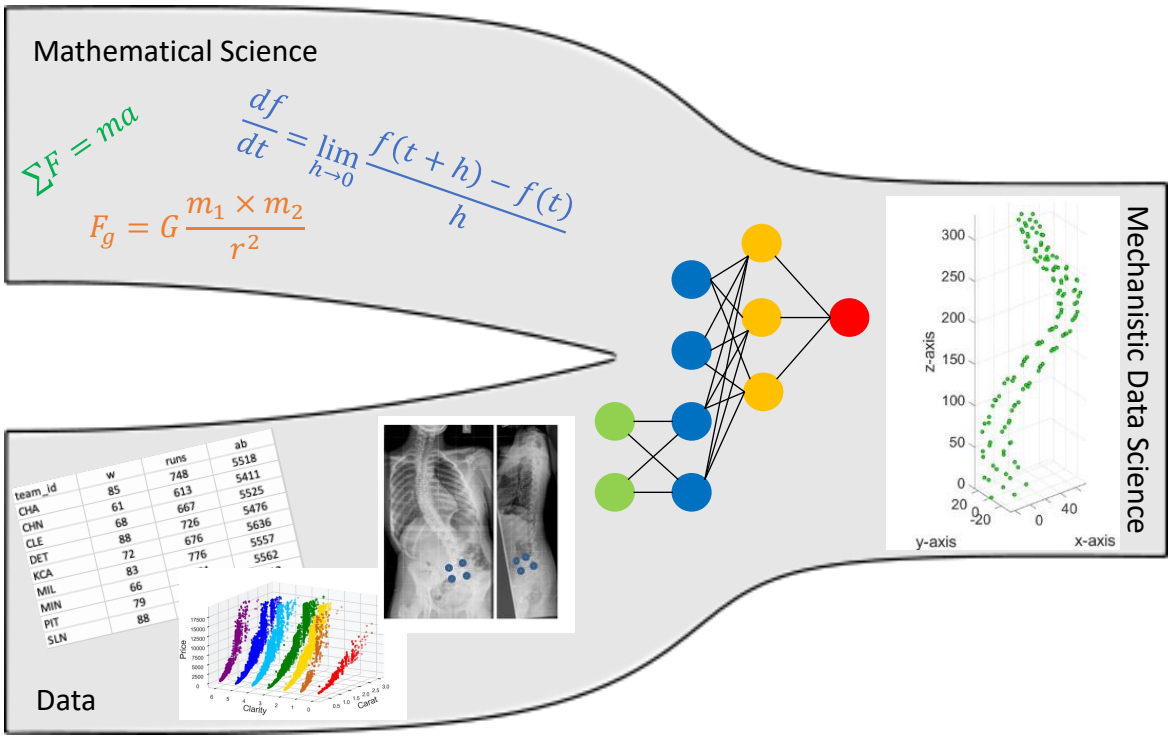


Figure 1 Schematic of Mechanistic Data Science.

1.1. A brief history of science: from reason to empiricism to mechanistic principles and data science

The history of science has been one of observations leading to theories leading to new technologies. In turn, the new technologies have enabled people to make new observations, which led to new theories and new technologies. This cycle has been repeated for thousands of years, sometimes at a slow pace, and other times at a very rapid pace.

Ancient philosophers such as Aristotle (384-322 BC) believed that scientific laws could be discerned through reason and logic. Aristotle reasoned that the natural state of an object was to be at rest and that heavier objects fell faster than lighter objects because there was more downward force on them. Aristotle believed in a geocentric universe (the earth is the center of the universe) and that the heavens were made of the quintessence, which was perfect and unalterable. In other words, there could be no supernovae or comets!!

These scientific ideas remained the benchmark for centuries until scientists in the Renaissance began questioning them. The astronomer Nicolaus Copernicus (1473-1543 AD) proposed a heliocentric model of the universe in which the sun was the center, not the earth. The idea that humans were not the center of the universe was very difficult for humans to accept. However, the publication of Copernicus' seminal work *On the Revolutions of the Celestial Spheres*, published in 1543 around the time of his death, set off the Copernican Revolution in science which resulted in a major paradigm shift away from the Ptolemaic geocentric model of the universe.

The work of Copernicus was later supported by the observations of Tycho Brahe (1546-1601 AD). Tycho was a talented astronomer who recorded many accurate measurements of the solar system which became the foundation for future astronomical theories. Tycho was followed by Johannes Kepler (1571-1630), who used the data collected by Tycho to formulate scientific laws of planetary motion which could predict their past or future position. Of particular note, he determined that the planets moved in an elliptical orbit around the sun, not a circular orbit, in contrast to the ancient Greeks, who thought the universe was geocentric and planetary motion was circular.

Galileo Galilei (1564-1642) is known as the father of the scientific method because of his systematic combination of experimental data and mathematics. He was a contemporary of Kepler who was the first scientist to use a telescope to observe celestial bodies and championed the heliocentric model of the universe. In 1638, he published *Discourses and Mathematical Demonstrations Relating to Two New Sciences* (or better known by the abbreviated name *The Two New Sciences*) in which he laid out the fundamentals of strength of materials and motion.

1.2 Galileo's study of falling objects

One of Galileo's major contributions was his study of motion and his ability to discern primary forces such as gravity from secondary forces such as friction and wind resistance. A notable example was Galileo's study of falling objects. We have all seen a dense, weighty object like a baseball fall faster than a lightweight object like a feather or a piece of paper. As discussed previously, ancient philosophers such as Aristotle had postulated that heavier objects fall faster than lighter objects in proportion to their mass. This remained the generally accepted theory of gravity until Galileo began studying falling objects in the late 1500's. Around 1590, when Galileo was a professor of mathematics at the University of Pisa in Italy, he reportedly conducted

experiments (according to his student Vincenzo Viviani) by dropping objects of different masses from the leaning tower of Pisa to demonstrate that they would fall at the same speed.[1]

Many years later in 1971, astronaut David Scott performed a similar experiment on the moon in which he dropped a feather and a hammer at the same time. Because the moon has almost no atmosphere (and thus no air resistance to slow the feather), the feather and the hammer hit the ground at the same time.[2]

1.3 Newton's laws of motion

Isaac Newton (1642-1726) was a scientist and mathematician best known for his laws of motion and the invention of calculus. Newton's laws of motion are a classic example of a law that was developed through the scientific method. He synthesized many years, decades, and centuries of observations, experimental data, and theories by scientists and mathematicians such as Galileo, Kepler, and Copernicus, into a new understanding of motion. In 1687, Newton published his work *Philosophiae Naturalis Principia Mathematica* (better known by its abbreviated name *Principia*), which has become one of the most classic scientific texts in history. In this book, Newton laid out three fundamental laws of motion:

1. Law of inertia – an object in motion tends to stay in motion and an object at rest tends to stay at rest unless some force is applied to it.
2. Law of force balance – changing the motion requires a force to be applied, which leads to the classic equation Force = mass * acceleration.
3. Law of reaction forces – for every action/force, there is an equal and opposite reaction.

These three seemingly simple laws of motion account for and describe nearly all the motion that we see and experience in the world around us even to this day. In fact, it was not until the early 20th century when Albert Einstein's theory of relativity was needed to describe the motion of objects traveling at a significant fraction of the speed of light.

Since the time of Newton, tremendous technological strides have been made by coupling fundamentals laws of science with creativity to meet the needs of people. One emblematic example is Thomas Edison (1847-1931) and the light bulb. One of the most basic human needs is to see in the dark, something that was mostly accomplished by fire until the invention of the electric light bulb. Although Edison did not invent the first light bulb, he took it from a crude concept to a mainstream technology. Early light bulbs would only last around 14 hours, but Edison improved them so that they worked 1200 hours. This was accomplished through a long and arduous process of trial and error. When asked about the 1,000 failed attempts at inventing the lightbulb, Edison famously replied that he "didn't fail 1,000 times. The light bulb was an invention with 1,000 steps".

---

[1] "Galileo's Leaning Tower of Pisa experiment," *Wikipedia*. Aug. 27, 2020, Accessed: 8 Sep. 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Galileo%27s_Leaning_Tower_of_Pisa_experiment&action=history.

[2] https://nssdc.gsfc.nasa.gov/planetary/lunar/apollo_15_feather_drop.html

This Edisonian *brute force* method is a purely empirical approach for applying science for technological development. This method involves pursuing and achieving a goal by building a design and testing it, and making small modifications based on the results of the previous tests. These steps are repeated until the inventor is satisfied with the design. During this process, the inventor is learning what works and what does not work, and this information can be used for calibration of parameters in conjunction with the applicable scientific laws. Another offshoot of the Edisonian style brute force method is that lots of data for the various trials is collected which can be very informative for future work. *The combination of the calibrated mechanistic principles and the data collected can be used to accelerate future development*.

One application of using collected data is artificial intelligence (AI) and neural networks in which data is used to guide decision making. An early success story for AI involved the chess matches between the IBM Deep Blue computer and chess champion Garry Kasparov. In 1985, Carnegie Mellon University began a project was started to "teach" a computer to play competitive chess. Over the next decade, the computer algorithm was trained using data from 4,000 different positions and 700,000 chess games by chess grandmasters. In 1996, Deep Blue actually won a single chess game in a six-game match against chess champion Garry Kasparov. In 1997 rematch, Deep Blue won the entire match against Garry Kasparov.

Implicit to the computer programming for playing games like chess is *game theory*, and no one is more synonymous with game theory than John Nash (1928-2015). Nash laid out his theory for achieving an optimal solution to non-cooperative games which came to be known as Nash equilibrium. It states that in a non-cooperative game with known strategies and rational players, the game achieves a state of equilibrium if no player can improve their position by unilaterally changing their strategy. Nash equilibrium is often illustrated by the prisoner's dilemma in which two prisoners apprehended together are interrogated in separate rooms. Both prisoners are given three choices: 1) freedom if they confess before the other prisoner but their partner receives extensive jail time, 2) minimal jail time if neither confesses, or 3) extensive jail time if they don't confess but the other prisoner does confess. It can be easily seen that each prisoner will achieve their own best outcome by confessing first rather than trying to cooperate with each other by not confessing. Since its discovery, Nash equilibrium has become one of the most important concepts in the game theory approach to artificial intelligence and decision making for neural networks.

The dawn of the new millennium also brought about the information age in which information and data are collected and categorized as never before in history. No more going to the library to look up information for a research paper – just Google it. The information age could also be called the data age because of the large amounts of data being collected. The challenge is to turn that data into information and using it synergistically with already known and established understanding of our world. In this book we have called this synergy Mechanistic Data Science.

Human progress has been greatly accelerated by our ability to understand and control the world around us. A large part of this is because of science and engineering. Since the times of Galileo and Newton, the fundamental principles of materials and motion have been further studied and formalized. Special technical fields of study have developed that feed and interact with each another in a symbiotic manner.

1.4 Science, Technology, Engineering and Mathematics (STEM)

Science provides us with a set of fundamental laws that describe nature and natural phenomena. From physics to chemistry to biology, once the natural phenomena are scientifically understood and described, we as humans are able to predict a particular outcome without testing for each possible outcome. Science is heavily reliant on mathematics for the "language" in which its laws are written and is also dependent on engineering and technology for applying scientific findings and developing new tools to enable future discoveries.

Mathematics is the unifying language of the physical sciences, and as such, the development of mathematics is integral to scientific progress. The understanding and capability of data scientists in mathematical topics such as algebra, geometry, trigonometry, matrix algebra, calculus foster the progress in science and engineering.

Engineering is the application of scientific principles for design and problem solving. In other words, we can make things. Engineers use data collected regarding the needs and wants of society and then use the principles of science, invoke human creativity and apply manufacturing craftsmanship to design, develop and produce products that address these societal needs and wants.

Technology is the implementation of products and capabilities developed through science and engineering. This generally takes the form of actual products on the market and in use today. Technology draws heavily on scientific discovery and engineering development to be able to address challenges, grow the economy, and improve efficiency.

The **scientific method** provides an organized methodology for studying nature and developing new scientific theories. A basic form the scientific method is:

- Observe: the subject of interest is studied and characterized from multiple standpoints. This first stage involves data collection in order which is organized and analyzed in order to move to the next stage.
- Hypothesize: based on the observations and the early data collected, a hypothesis (proposed explanation) is developed.
- Test: experiments are conducted to evaluate and challenge the hypothesis. In this stage, extensive amounts of data are collected and analyzed.
- Theory or law: if a hypothesis is not proven to be false during the testing challenges then it is established as a theory. Theories that are considered to be fundamental and widely accepted are often described as laws. Note that often times a theory or law is established with limitations for when it is valid (e.g., Newton's laws of motion are valid for speeds much less than the speed of light, but Einstein's theory of relativity is required for objects moving at speeds close to the speed of light.)

There are often many special cases for a scientific law. In the above-mentioned example of a falling object, it is necessary to account for air and wind resistance when comparing a falling hammer versus a feather. They will fall at the same rate in a vacuum or on the surface of the moon, but in normal atmospheric conditions, the effect of air resistance makes a noticeable

difference in the rate of falling. In this case, the law of gravity still applies, but it must be coupled with data on wind resistance, object density and aerodynamics, in order to describe the effect of wind resistance.

Once a scientific law has been established, engineers can use this information to make design calculations in the continuing quest to design and build new products. For example, understand gravity and falling objects allows engineers to design and build objects that can fly, whether they are a backyard water bottle rocket as shown in the figure below, or a high tech reusable rocket like the SpaceX rocket that can return components to Earth and land upright on a barge in the ocean.

Figure 2 (a)backyard water bottle and baking soda rocket (b)SpaceX Falcon Heavy rocket launch (Reuters/Thom Baur).

1.5 Data Science revolution

Recent years have seen a revolution in data science as large amounts of data have been collected on a vast array of topics. For instance, the ubiquitous smart phone is constantly collecting and transmitting info about grocery store purchases, travel routes, and search history. Companies like Facebook, Google, and Instagram collect and utilize data posted on their site for various marketing purposes such as targeted advertisements and purchase recommendations. These sites match demographic information such as age, gender, and race with internet browsing history, purchases made, and photos and comments posted to predict future behavior, such as whether you are likely to buy a car or make some other large purchase. These predictions can be used to target advertisements at the right audience.

Data science has been heavily used for product development, from the concept stage to engineering and manufacturing to the customer. Data collected at each stage can be used to improve future products or identify the source of problems that arise. For instance, manufacturers regularly collect customer data to understand the “voice of the customer”, which is used to plan future product models. Furthermore, as a product is being manufactured, data is being collected at every step of the manufacturing process for process control. The increased use of sensors has given rise to the internet of things (IOT) in which data is automatically collected and transmitted over the internet for analysis without requiring explicit human interaction.

1.6 Data Science for Fatigue Fracture Analysis

Disasters have often been a driving force for exploration of new areas or to develop a much deeper understanding of existing areas. The data, techniques, and methods generated from these explorations in turn becomes a boon for engineers and product designers when designing new products or working to improve existing designs. One such area of scientific and engineering exploration is the study of fractures and failures due to fatigue. Fatigue is the initiation and slow propagation of small cracks into larger cracks under repeated cyclic loading. The formed cracks will continue to get longer and longer while the product is being used under normal operating loads until they become so large that the structure fails catastrophically.

Consequences of fatigue: On April 28, 1988, Aloha Airlines flight 243 took off from Hilo, HI on a routine flight to Honolulu, HI. The Boeing 737 airplane had just reached cruising altitude when a large section of the separated from the plane (see Figure 3). The pilot was able to successfully land the plane on the island of Maui, although one flight attendant was lost in the incident. Post-incident inspection of the airplane showed that small cracks had initiated from the rivet holes which were used to were used to join the separate pieces of the airplane fuselage. The cracks had propagated slowly from hole to hole until the resulting crack was sufficiently large that the structure could no longer support the service loads in the forward section of the plane. A commercial airplane is pressurized for every flight, which stresses the fuselage, in addition to takeoff and landing loads, and vibration loads. The subject airplane had accumulated 89,680 flight cycles and 35,496 flight hours prior to the incident.[3]

It should be noted that this incident resulted in the formation of the Center for Quality Engineering and Failure Prevention led by Prof. Jan Achenbach at Northwestern University and with which two of the authors collaborated with extensively in the past. Prof. Achenbach received both the National Medal of Technology and the National of Science partly for his important scientific work on the non-destructive detection of fatigue cracks.

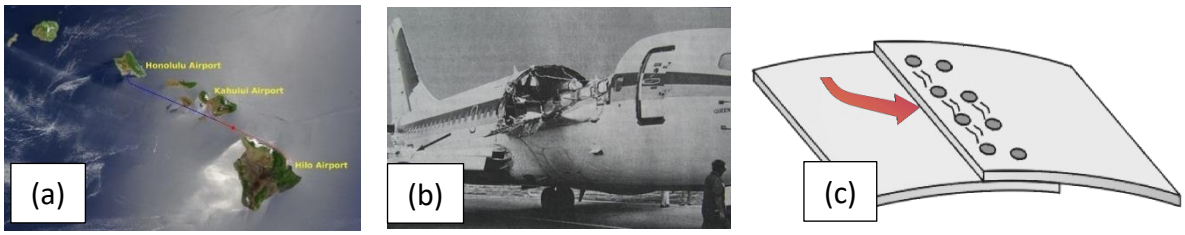


Figure 3 Aloha Airlines flight 243 fatigue fracture example (a) scheduled and actual flight paths of flight 243 (b) Boeing 737 airplane after fuselage separation (c) schematic illustration of fatigue crack growth between rivet holes.

Fatigue design methodology: One of the most common methods for designing and analyzing for fatigue is a mechanistic data driven methodology known as the stress life method. In this methodology, the material of the design has been tested at many different stress levels to determine how many load cycles to failure. That stress amplitude is then plotted on a graph vs. the number of cycles to failure for that material. This plot is commonly referred to as an S-N

[3] "Aloha Airlines Flight 243," *Wikipedia*. Sep. 06, 2020, Accessed: 8 Sep. 2020. https://en.wikipedia.org/wiki/Aloha_Airlines_Flight_243

curve. When used for design analysis, the cyclic stress amplitude at a location of interest is either measured by laboratory or field testing or computed using finite element analysis or hand calculations. Using the computed or measured stress amplitude and the appropriate S-N curve, the fatigue life can be estimated.

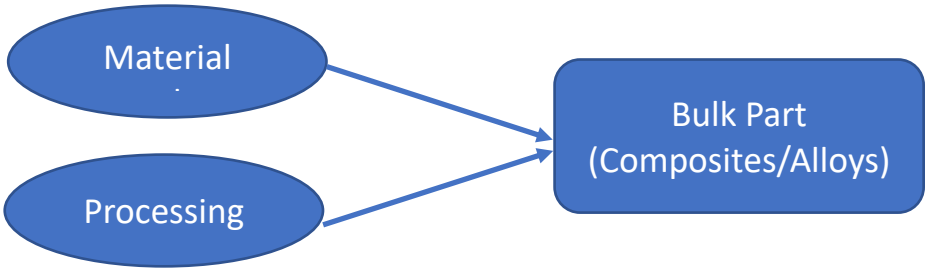
There are many factors involved in the initiation and propagation of fatigue cracks, such as material strength, microscopic impurities and voids, and surface roughness. For each material of interest, many controlled laboratory tests are conducted on standardized specimens at a range of stress levels to measure how many cycles the material can endure before fracturing. This data-driven approach has been necessary because of the relatively large number randomly occurring of factors that are involved.

Fatigue cracks generally initiate at or near the surface of the material. It has been found that parts with rougher surfaces will have shorter fatigue lives, with the effect being more pronounced at higher fatigue lives. For many materials, a large amount of testing has been performed to characterize the effect of the surface roughness due to the manufacturing process used to make the part (eg. polished surface, machined surface, or as-cast surface finish).

1.7 Data Science for Materials Design: "What's in the cake mix"

The macrostructure (or physical structure you can see and hold in your hand) is composed of trillions of atoms of different elements which are mixed and organized in a certain way at the small-scale sub-structure, and this organization depends on the particular material being used. This small-scale sub-structure is referred to as the microstructure if a powerful microscope is required to see it and is called the mesoscale if you need to use a magnifying glass or just look very closely.

The overall macrostructural performance of a bulk part is controlled in large part by the microstructure of the material used to make the part. For example, if a pastry chef is baking a cake, the flavor, texture and crumble of the cake are controlled by the ingredients, the mixing, and the baking time and temperature. One can think of the microstructure as "what's in the cake mix".



"Microstructural" ingredients "Macrostructure"

Figure 4 "Cake Mix" material microstructure example ("Classic Carrot Cake with Cream Cheese Frosting." Once upon a chef with Jenn Segal. https://www.onceuponachef.com/recipes/carrot-cake.html.)

Engineered components are often evaluated for strength, stiffness, and fracture resistance (as opposed to taste and texture for a cake). A simple example is to consider an ice cube. If one

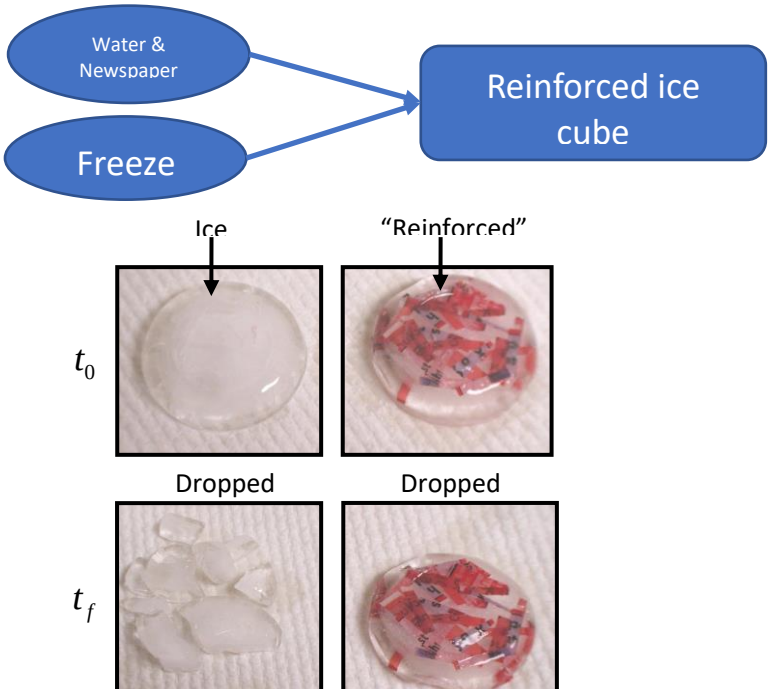


Figure 5: Ice with and without reinforcement dropped to show effects of reinforcement on fracture resistance. Experiment by Northwestern University Prof. Yip Wah Chung, 2003.

were to make an ice cube from only water, the ice cube would likely shatter into many pieces when dropped on a rigid surface from a sufficient height. However, if other ingredients were added to the water when the ice was made, the resulting ice cube would likely be more resistance to shattering, depending on what was added. For example, if strips of newspaper were added to the ice, the resulting reinforced ice cube would not shatter when dropped from the same height as the unreinforced ice cube. This is because the mesostructure of the newspaper in the ice increases the toughness of the ice cube and resists cracks from propagating through the cube as it impacts the rigid surface.

If one realizes that we have entered the digital age and newspaper is no longer readily available, then new composite material needs to be developed to replace newspaper filler. Given below is a sampling of alternative reinforcement materials that could be used to make a composite cube structure. The evaluate the impact fracture resistance of the various cubes, several composite ice cubes were made and dropped 14 feet (impact speed of 30 ft/sec) onto a concrete surface. Results showed that the addition of wood chips was most effective in preventing fracture of the cubes on impact.

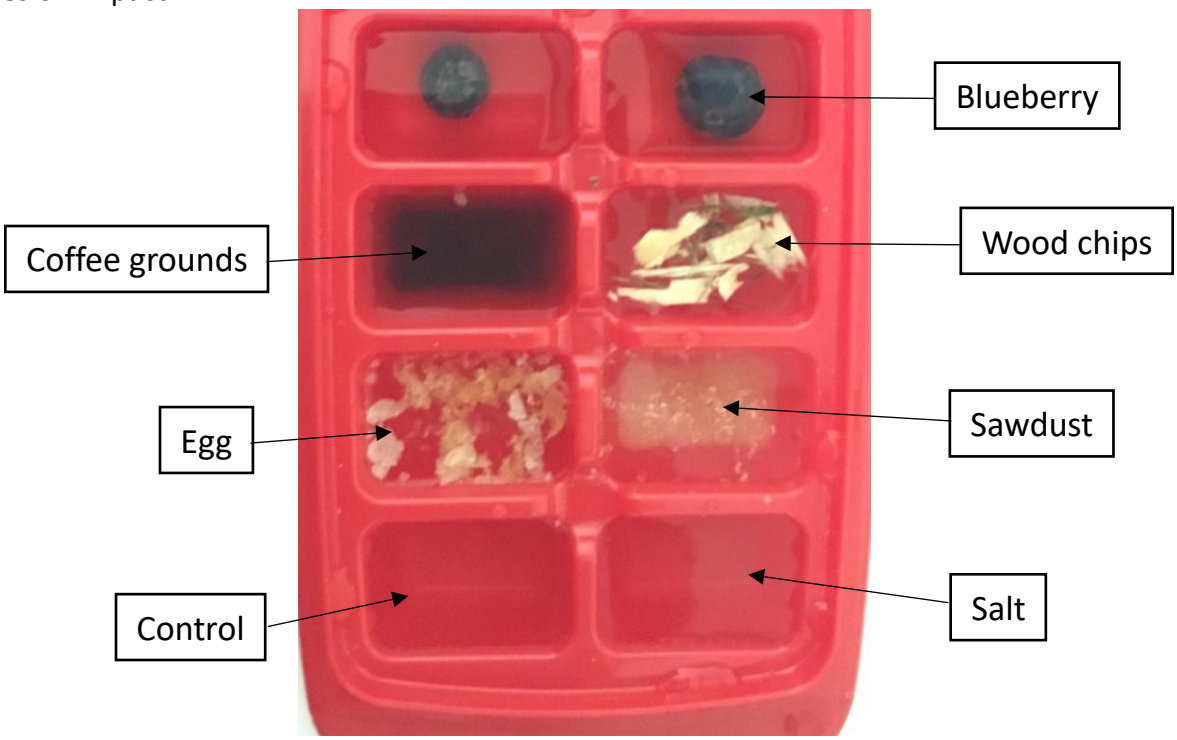


Figure 6: Composite ice cube experiment by Northwestern University Prof. Mark Fleming and Carmen Fleming, 2020.

Table 1 Drop test results for ice cubes formed with various reinforcement materials.

Filler material	Result
Control	Fractured, big chunks
Salt water	Small chunks
Egg	Fractured, big chunks
Sawdust	Fractured, small chunks
Wood chips	Very little fracturing
Coffee grounds	Fractured, big chunks
Blueberry	Fractured, big chunks

From everyday applications to materials design

The reinforced ice cube can be considered as a metaphor for an engineered composite material in which the mesoscale structure enhances the overall structural performance. Consider the material substitutions shown in Figure 7. A composite material in its most basic form consists of a matrix material with some reinforcing material embedded inside. If the matrix ice material is replaced with an epoxy and the newspaper is replaced with carbon fibers, a carbon fiber reinforced composite material results. On the other hand, if the matrix ice material is replaced with rubber and the newspaper is replaced with steel or polymer cords, a tire can be made.

An example of MDS Framework for next generation tire tread material design

One of the fundamental questions that tire industry faces is on the durability of the tire. The unpredictable weather conditions and road surface conditions that each tire faces every day have significant impact on its durability. One of the key materials property metrics that can be related with the tire material performance is the tan(δ). For tire material it is desirable that it has high tan(δ) for low temperature (provide better ice and wet grip) and low tan(δ) at high temperature (provide better rolling friction). It is noteworthy to mention that approximately 5-15% of the fuel consumed by a typical car is used to overcome the rolling friction of the tire on the road. Therefore, controlling the rolling friction of tires is a feasible way to save **energy** (by reducing fuel consumption) and **environment** (by reducing carbon emission). We can also ensure the **safety operation** of tire providing sufficient ice or wet grip.

The key performance metric tan(δ) is a function of the matrix materials, microstructure, and the operating conditions such as temperature and frequency. It is well known that adding filler improves the tire materials performance. But what fillers and their distribution to achieve optimized properties and performance is still an important research question. The design space combining different rubber matrices and fillers, microstructure and operating conditions can be enormous that experimental or simulation technique is not feasible at all. The mechanistic data

science approach can provide an effective solution to explore the design space leveraging the data science tools through revealing the mechanisms and construction of necessary accurate and efficient reduce order surrogate model. This approach will enable the industrial practitioner to perform rapid design iteration and expedite the decision-making process.

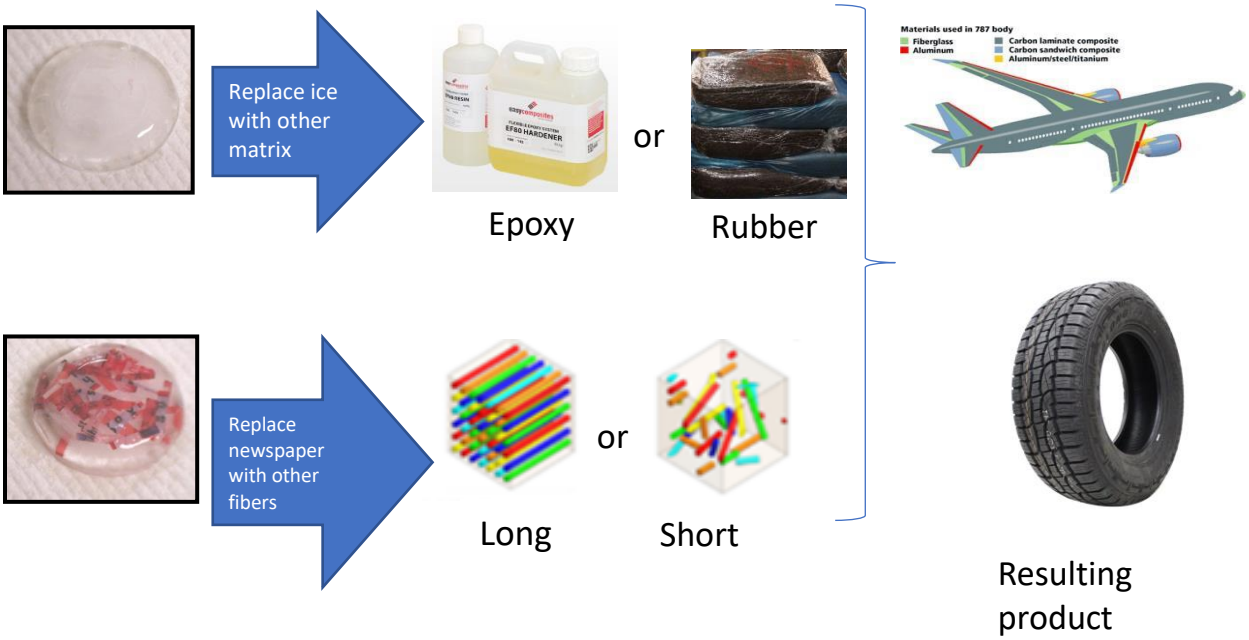


Figure 7 Ice cube to engineered materials analogy for materials design.

Gold and gold alloys for wedding cakes and wedding rings

Pure gold is so ductile that it can be rolled into sheets that are so thin that it can be used to decorate cakes and subsequently eaten. While this expensive application of pure gold is an interesting possibility for a wedding cake, gold is more often used for jewelry such as a wedding ring. As such, it needs to hold a specific shape. Materials like gold are strongly influenced by the microstructure of the material. Pure 24K gold is extremely ductile and malleable, meaning that it can be reshaped by rolling and pounding without cracking. Gold jewelry is generally made from 18K or 14K gold, which is strengthened by alloying (mixing with other elements) it with other metals. As an example, 18K gold contains a mix of 75% pure gold, 10% copper, 8% nickel, 4.5% zinc and 2.5% silver.

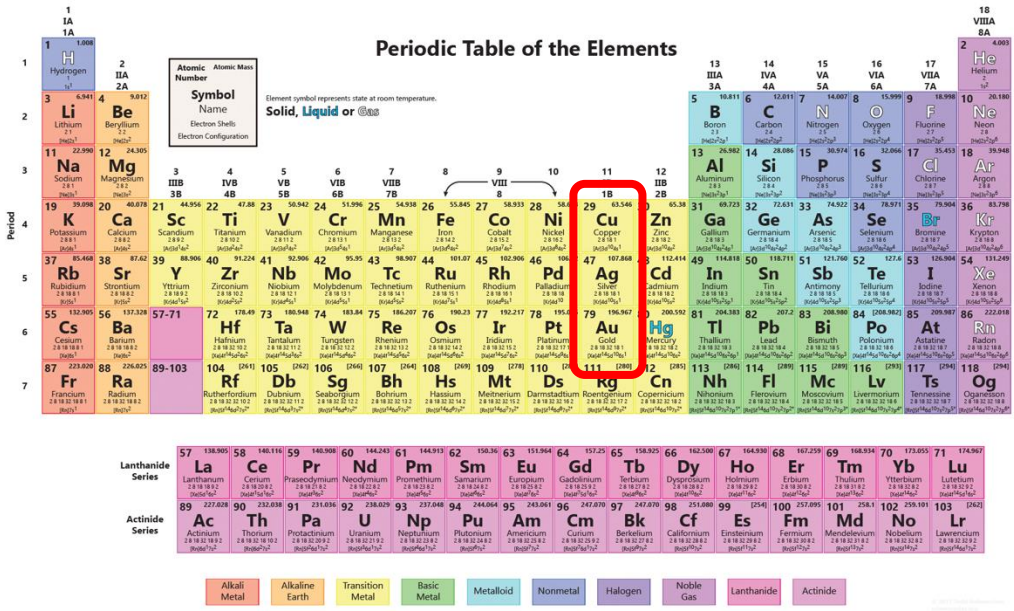


Figure 8: Periodic table of elements (Source: Sciencenotes.org). The red box signifies some of the elements used for making gold alloys.

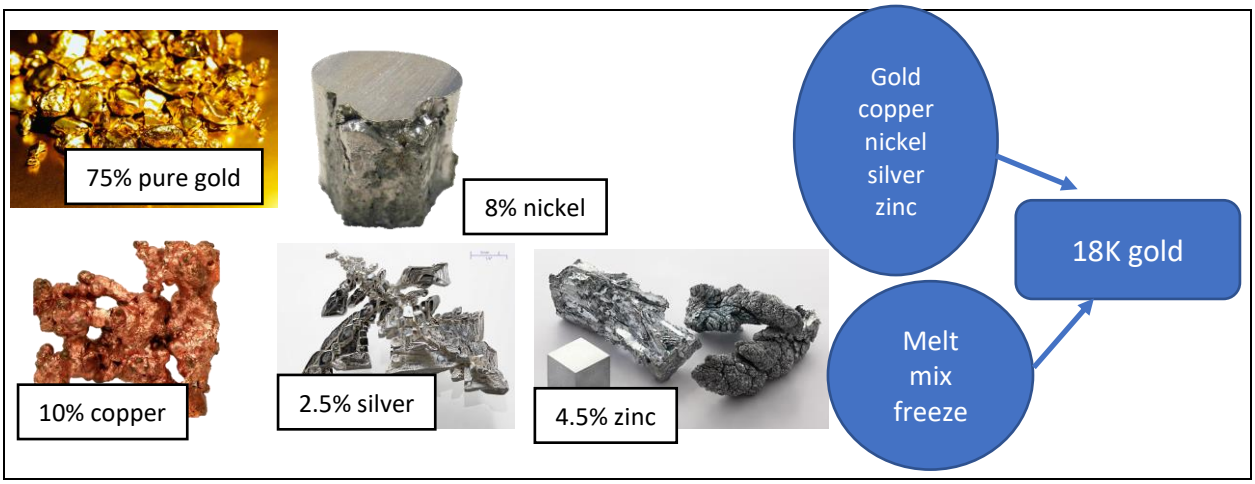


Figure 9: Elements making up 18K gold, which is a common gold alloy for jewelry.

1.8 Twenty-first century data science

AlphaGo

One very interesting data science accomplishment is the use of deep learning neural networks for the complex game of Go. Go is a complex strategy game created 3000 years ago in China. The game is played with black and white stones that are placed on a grid-filled board in an attempt to surround an opponents stones and to strategically occupy space. The possibilities of the game are astronomical, with 10^{170} possible configurations, which makes it much more

complex than chess. Until recently, the best Go computer programs could only play at a relatively novice level.

In 2014, a company called Deep Mind began working on a project that led to a program called AlphaGo and trained it to play using deep learning neural networks. It was initially trained with many amateur games and against itself, all the time improving its ability.

In 2015, AlphaGo played a match against reigning European champion, Mr. Fan Hui, and won a game. Then in 2016, AlphaGo beat 18-time world champion Mr. Lee Sedol. Since that time, additional versions have been released, including AlphaZero, which is capable of learning other games such as chess and shogi.

3D printing: from gold jewelry to customized implants

Recently, a new method of manufacturing structural parts called additive manufacturing (AM) or 3D printing has become popular. A report from the National Academy of Engineering identified 3D printing as a revolutionary new manufacturing technology capable of making complex shapes, and possibly one day printing new body parts.[4] Additive manufacturing generally works by building a part through depositing a thin layers of material one after the other until a complete part is formed. This process is able to form very complicated shapes, and has been widely used in many industrial applications from precious metals such as gold to customized 3D spinal implants (see Figure 10). To highlight some of the positive impacts that 3D printing is having around the globe, The top ways that the technology is making a difference for the environment, health, culture and more include disaster relief, affordable housing, more efficient transportation (less pollution), better and more affordable healthcare, 3D bioprinted organs, cultural and archeological preservation, accessible medical and lab devices, and STEM education.

There are a lot of data during 3D printing process and part qualification, such as huge amount of process parameters, spatical and temporal physical fields, microstructure, and mechaical properties and performances. Data science is extrmely useful for 3D printed part qualification and optimization for stablized printing processes and improved material properties.

[4] NAE Report on Making a World of Difference-Engineering Ideas Into Reality.

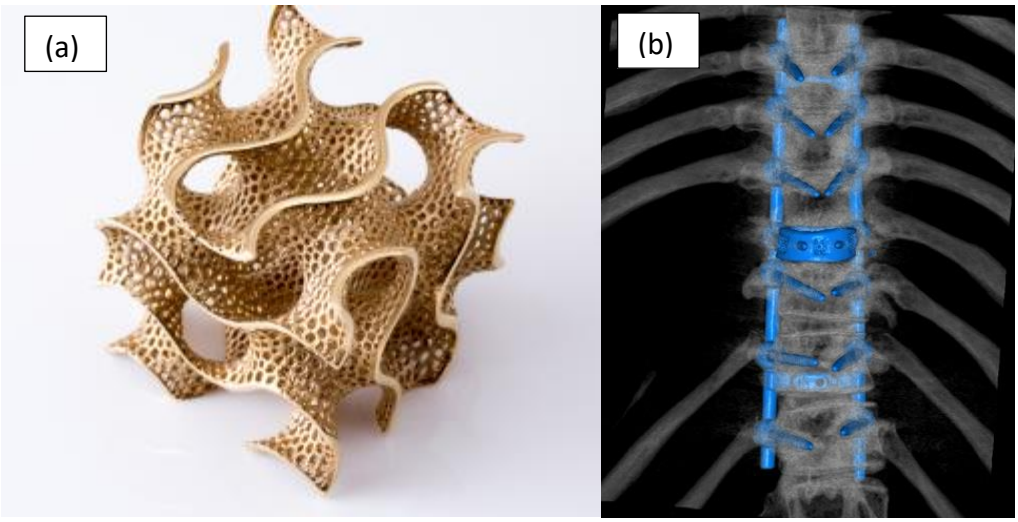


Figure 10 (a) A example of 3D printed gold jewelry.[5] (b) 3D custom implant to reconstruct a vertebra destroyed by a spinal tumor.[6]

1.9 Outline of Mechanistic Data Science Methodology

As shown in Figure 11, data science for solving engineering problems can be broken down into six modules, ranging from acquiring and gathering data to processing the data to making calculations.

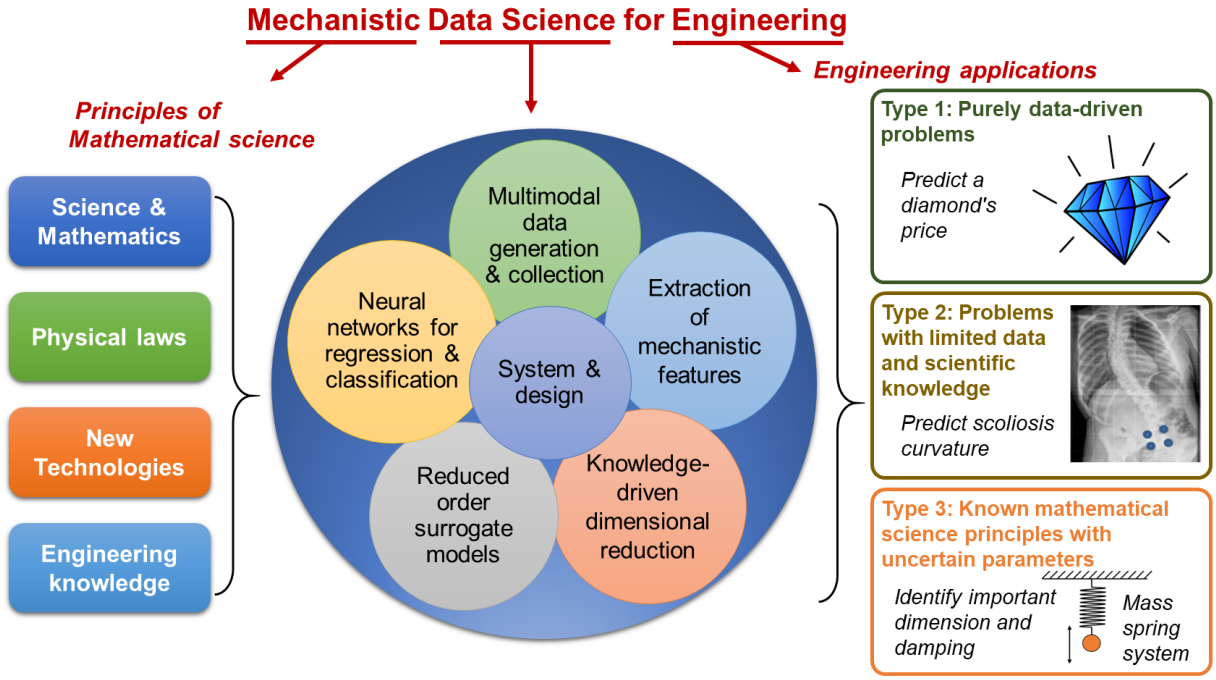


Figure 11: Schematic of Mechanistic Data Science for Engineering. Scientific knowledge is combined with data science to effect engineering design with the goal of obtaining knowledge for improved decision making.

[5] https://3dprintingindustry.com/news/eos-cooksongold-put-bling-3d-printing-precious-metal-printer-33041/
[6] https://www.spineuniverse.com/resource-center/spinal-cancer/3d-spinal-implants-glimpse-future

Mechanistic data science is the structured use of data combined with the core understanding of physical phenomena to analyze and solve problems, with the end goal of decision-making. The problems to be solved range from

Type 1 *purely data-driven*: problems with abundant data but undeveloped or unavailable fundamental principles. This type of problem can be illustrated through using data for the features of diamonds to determine the price. There is not an explicit "theory" associated with diamond prices. Instead, the price is determined by the complex interplay of many features such as size and sparkle.

Type 2 *limited data and scientific knowledge*: problems in which neither the data nor the scientific principles provide a complete solution. This type of problem can be illustrated the analysis of scoliosis patients. X-rays provides some data for the progression of spine growth but combining that data with finite element surrogate models in a neural network provides a good estimate of scoliosis progression.

Type 3 *known mathematical science principles with uncertain parameters*: problems which can be computationally burdensome to solve. This type of problem can be illustrated through a spring-mass example. Physics models of spring-mass systems typically assume a point mass, a massless spring, and no damping. Data collected on an actual spring-mass system can illustrate how to use data science to identify key physical factors such as damping coefficient directly from high-dimensional noisy data collected from experimental observations.

In this book, mechanistic data science will be broken into six modules:

Module 1: Multimodal data generation and collection

Large quantities of data are collected related to the topic under study. Multimodal data is data from various types of sources, such as different type of measuring instruments and techniques, models, and experimental setup.

Multimodal data generation and collection will be described in chapter 2 (subsections AAA are for general readers and subsection YYY are for advanced readers).

Module 2: Extraction of mechanistic features

Mechanistic features are the key pieces of data that will be used for further data science analysis. They often have to be computed from the raw data collected. It should be noted that data scientists generally describe the feature extraction process as the step where they spend a majority of their time.

The extraction of mechanistic features will be described in chapter 3 (subsections AAA are for general readers and subsections YYY are for advanced readers).

Module 3: Knowledge-driven dimensional reduction

Dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables, generally based on the mechanistic features extracted. One example is to cluster the data into a reduced set of data.

Dimension reduction will be described in chapter 4 (subsections AAA are for general readers and subsections YYY are for advanced readers).

Module 4: Reduced order surrogate models

Reduced order surrogate models reduce the computational complexity of mathematical models in numerical simulations, generally based on fundamental or physical principles. For example, mathematical or physics models can be developed around the clustered data from the dimension reduction process.

Reduced order surrogate modeling will be described in chapter 5 (subsections AAA are for general readers and subsections YYY are for advanced readers).

Module 5: Neural networks for regression and classification

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

Regression and classification will be described in chapter 6 (subsections AAA are for general readers and subsections YYY are for advanced readers).

Module 6: System and Design

System and design tie the other modules together. The data science is coupled with the mechanistic principles in order to complete an analysis and make decisions.

System and design will be described in chapter 7 (subsections AAA are for general readers and subsections YYY are for advanced readers).

1.10 Examples describing the three types of MDS problems

Examples are provided in this section to illustrate the application of mechanistic data science for the three types of problems describes in Section 1.9. Examples ranging from strictly data-centric problems to medical diagnosis to physics-based problems are provided to illustrate the broad applications of MDS.

Determining price of a diamond based on features (pure data science – Type 1)

Jim and Maddie are two young people who want to get married. Jim wants to give Maddie a traditional diamond ring that really sparkles. However, they are recently graduated from college and have student loans to pay off, as well as many other expenses. Jim realizes that he needs to study what makes diamonds sparkle in order to get the best diamond ring that he can afford. Jim took a mechanistic data science course in college and decided to use that analytical capability when studying the features and prices of diamonds.

A cursory study of diamonds shows that they have some very impressive properties and can be used for industrial applications as well as jewelry. Diamonds are the hardest substance on earth, which make them popular for surface coatings in which wear resistance is important, such as cutting and drilling tools.

Diamonds for jewelry are known for their sparkle and impressiveness, which are features that are not easily quantified. However, they are functions of other features that can be quantified. The classic, best-known, quantifiable features of diamonds are the 4-C's (Cut, Clarity, Color, Carat), but other features such as depth and dimension are also reported. A combination of all these features is used when determining the price of a diamond.

Multimodal data collection: Jim found a large repository of data on diamond features and prices can be found at www.kaggle.com, which is a subsidiary of Google. For his analysis, a database of 53,940 diamonds is catalogued with information for ten different features along with the price for each diamond. A sample of some of the data is shown in the table below:

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31

Feature Engineering: The diamond data available on Kaggle is a good first start, but Jim needed to do some feature engineering, or initial data processing, to get it into an appropriate form for mathematical analysis. This includes removing missing values and converting alphabetic and alphanumeric scores to numerical scores (e.g., the cut of a diamond is rated as Excellent, Very Good, Good, Fair, and Poor. These ratings are converted to 1, 2, 3, 4, and 5.). In addition, it is often useful to normalize the data.

Dimension reduction: Jim found that for the analysis he was performing, certain features were often more useful than others. Since he was interested in how a diamond sparkles, the clarity and color were more interesting than the geometric features such as size distribution.

For his analysis, Jim created a correlation matrix to help separate the relevant from the irrelevant features. From this, Jim selected four features for further analysis: carat, clarity, color, y.

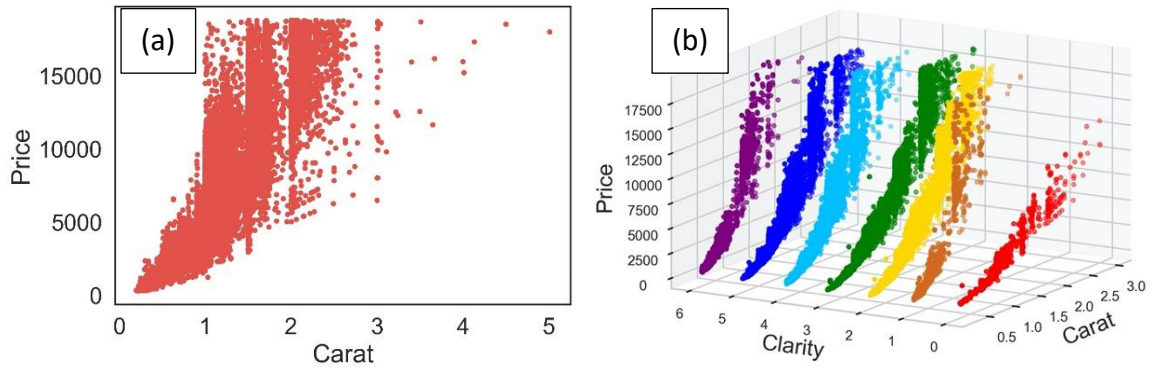


Figure 12 Diamond price vs carat (a) parameters combined (b) separated by clarity.

Regression and classification: Regression is the process of developing a mathematical relationship between variables in a dataset. Jim first performed the most basic form called **linear regression**, which tries to determine where to draw a line through the middle of the data. In the plot of the price vs. carat data shown above, the black line represents a linear regression best between the price and the carat of the diamonds. The **correlation** of the data describes how close the actual data is the regression line. It can be seen from looking at the graph that when all the diamond features are lumped together, there is a lot of scatter of the red data points around the black line, resulting in a low correlation. To achieve a better fit, Jim needed to consider additional features. The graph on the right shows the price vs. carat data subdivided by clarity. It can be seen that the price per carat increases more quickly for higher clarity diamonds.

Multivariable linear regression. Jim then decided to perform multivariate linear regression in order to consider the effects of multiple features simultaneously. He found that the correlation of the regression improved as more features are considered. In the graphs below, the prediction using a linear regression is plotted versus the actual data. It can be seen that as the number of features considered increases that the amount of scatter decreases.

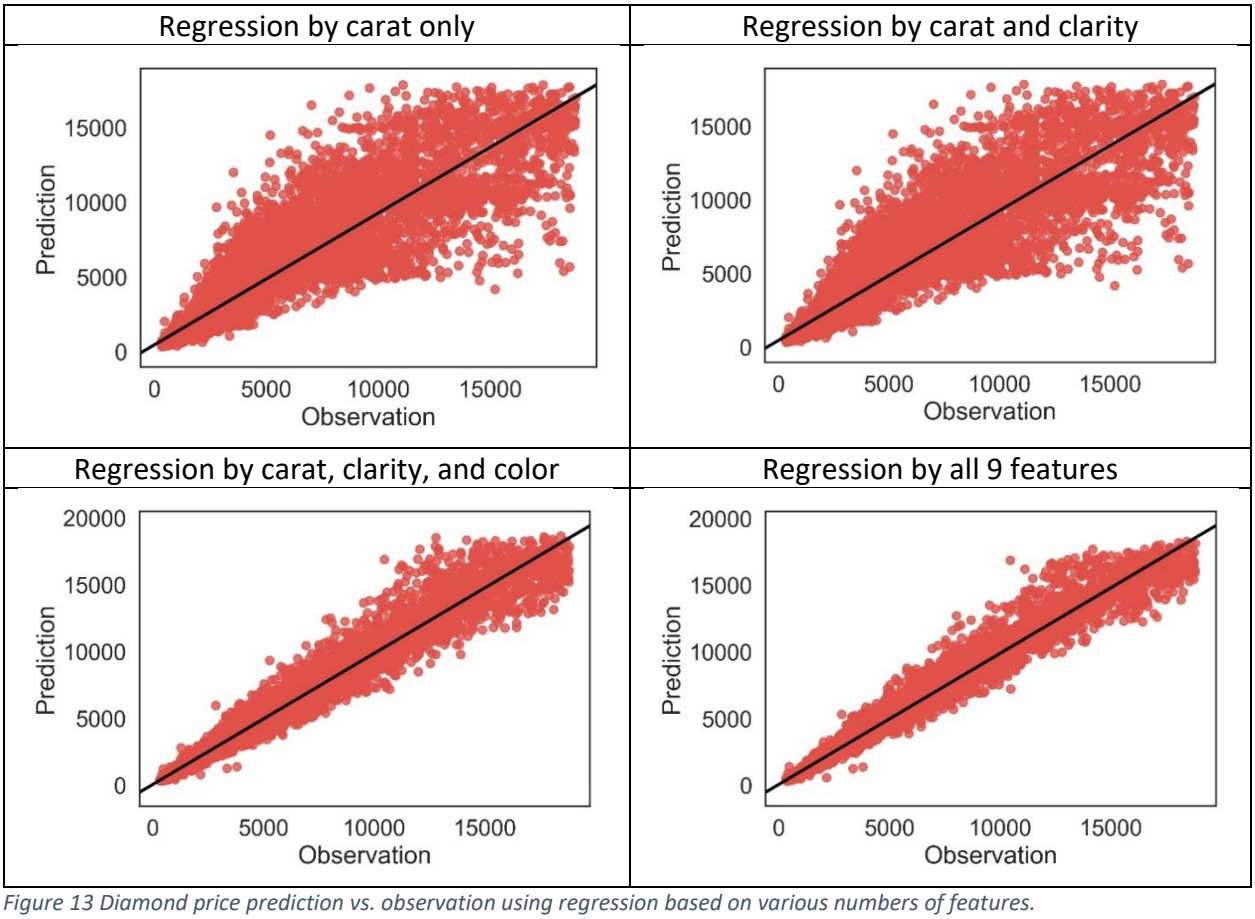


Figure 13 Diamond price prediction vs. observation using regression based on various numbers of features.

Once the multivariate linear regression is complete, Jim was able to estimate the price of a diamond using multiple features, and the nonlinear nature of diamond pricing is seen. For

instance, using the regression data, the following diamond prices are found for different sized diamonds:

1 carat diamond è $4,600

2 carat diamond è $17,500

In short, Jim found that when it comes to diamonds $1 + 1 \neq 2$.

Jim chose …

Predicting patient-specific scoliosis curvature (mixed data science & surrogate – Type 2)

Mechanistic data science can be used to analyze the progression of Adolescent Idiopathic Scoliosis (AIS), and someday soon provide a way to virtually assess the effectiveness patient-specific treatments before starting the actual treatment. AIS is a condition in which the adolescent spine curves in an unnatural manner. Recently, mechanistic data science has been used to study the progression of this condition.

The analysis and diagnosis of AIS begins with medical imaging of the spine. Two types of images are used for the analysis – X-rays and magnetic resonance imaging (MRI). X-rays of a patient are taken from the front or anteroposterior (AP) view and the side or lateral (LAT) view to capture the position of the vertebrae that make up the spine (see Figure 14). The X-rays are repeated to document the progression of the scoliosis condition over time. An outline of each vertebra can be extracted from the 2D X-ray data. The 2D data points are projected to 3D data.

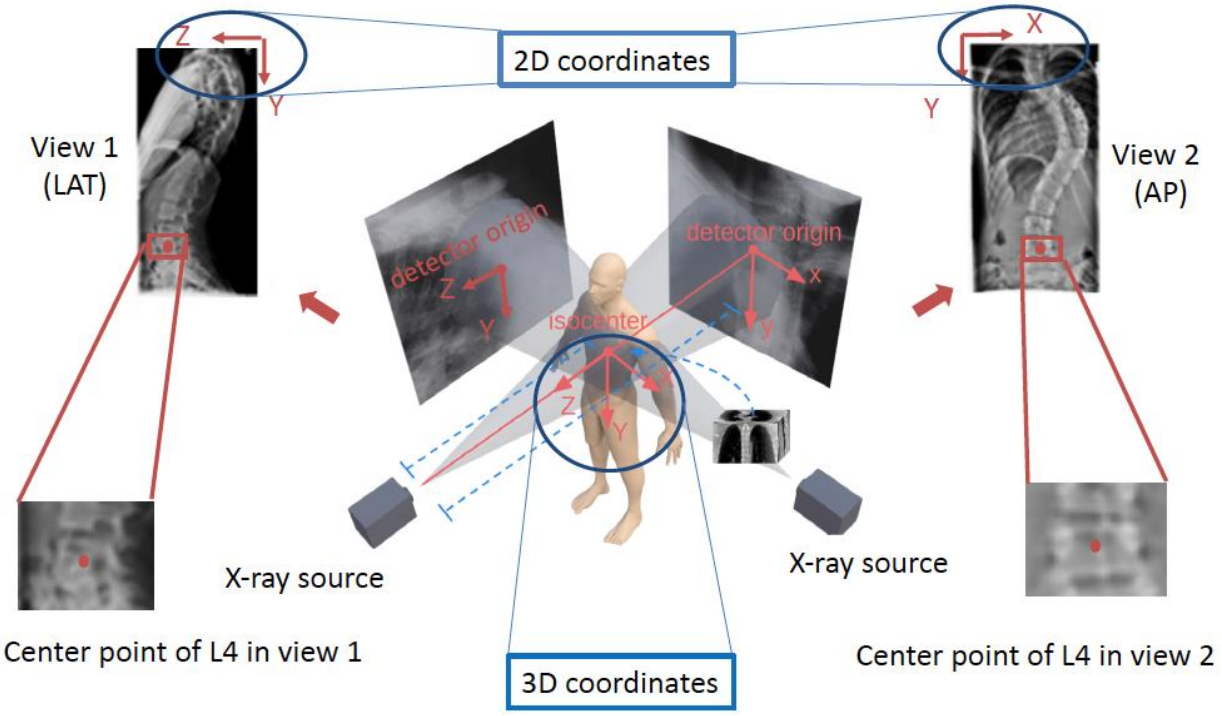


Figure 14 X-ray projections for collecting data on scoliosis progression.

In addition to X-rays, MRI's taken of a few patients provide a detailed 3D image of the entire spine. The MRI data from one patient's spinal vertebrae can then be used as reference or surrogate images for other patients. The surrogate model of the vertebrae can be adjusted to be patient-specific by combining it with data from the 2D X-rays.

The 2D X-ray data points for each vertebra that have been located in 3D space are overlaid on the surrogate model of the vertebra. For this analysis, the surrogate model of the vertebrae is taken from the MRI data of the spine, but there are other sources of vertebrae data that can also be used. Once the 3D X-ray data has been overlaid on the surrogate model of the vertebra, the surrogate model is scaled and adjusted to yield a patient specific model of the vertebra (see Figure 15). This process is repeated for each vertebra until the entire spine has been mapped for a particular patient by combining 2D X-ray images with generic models from a surrogate.

Key reference points, or landmarks, for each vertebra are also extracted from the 2D X-rays using image processing. The intersection of the two 2D projection is then used to locate the data points as 3D data points. (see Figure 15).

Using the 3D simplified model of the spine derived from the X-ray images, a detailed model of the spine was created using the atlas model vertebra. The generic surrogate model of each vertebra is updated based on the actual vertebra size and shape as shown in Figure 15 below. These more detailed vertebrae are assembled to form a patient-specific spine model.

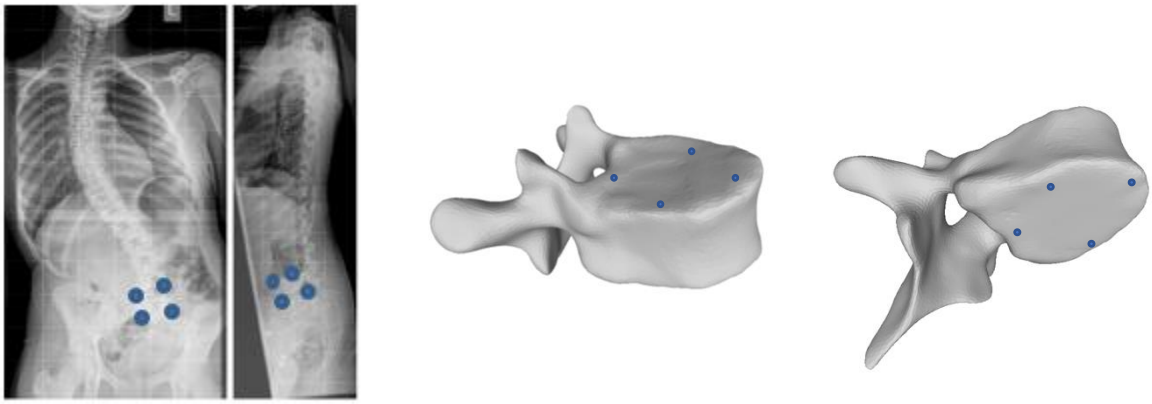


Figure 15: Surrogate geometry of a vertebra. The vertebra on the left is before being adjusted by the collected data. The vertebra on the right has been adjusted through the collected data.

The patient specific geometry of the spine can be used to generate a finite element model to compute the pressure distribution on each vertebra due to scoliosis. According to the Hueter-Volkmann (HV) principle, areas of a vertebra with higher stress grow more and areas with lower stress grow less. The pressure distribution computed using finite element analysis and the geometries of the vertebrae are updated based on the HV principle. The gravity load and the material properties of the spine material are also updated over time to reflect the changes in a specific patient due to aging. The computed stress results can be used as the input to a neural network, along with other factors such as the landmarks, the global angles, and the patient age, to predict how the spine would move over time. At this time, not enough information is known about modeling the materials of the growing spine and it is not possible to measure the pressure

distribution of the vertebrae pressing on each other. However, the results below show that through a combination of finite element computer simulation to compute the pressure distribution on the vertebrae and the data from the patient X-rays, it is possible to accurately predict the progression of scoliosis.

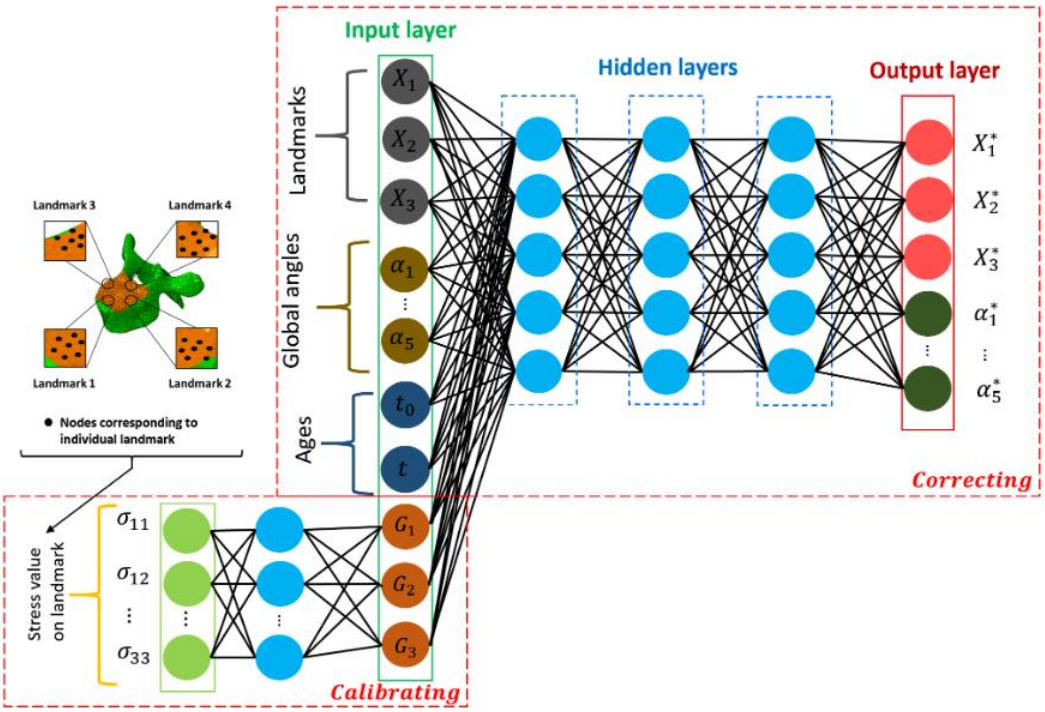


Figure 16 Neural network combining mechanistic models with X-ray data to predict spine growth in scoliosis patients.

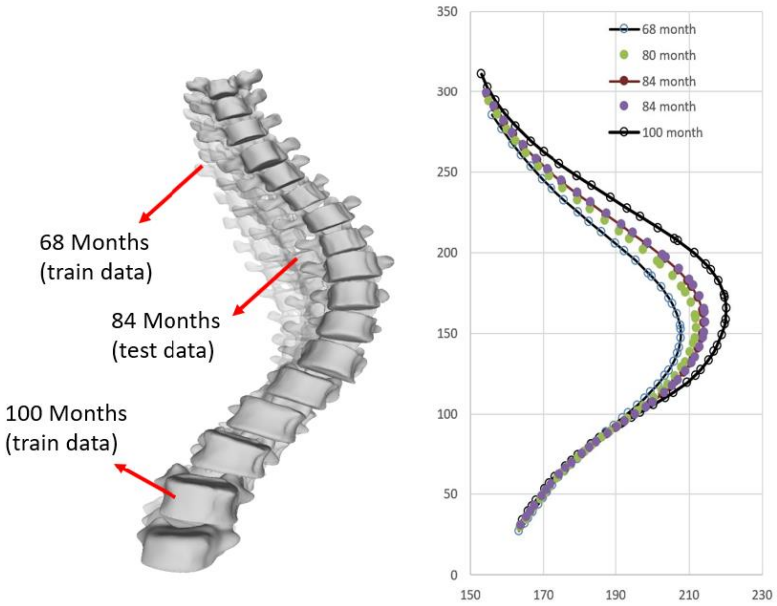


Figure 17 Predicted progression of spine growth in scoliosis patients

Identifying important dimension and damping in a mass-spring system (Type 3 problem)

A young engineer is given the job of running experiments with physical components and then analyzing the data. Unfortunately, the test data appears does not always clearly match with the theory learned in school. This is not a trivial problem, but rather a fundamental challenge in empirical science. Examples abound from complex systems such as neuroscience, web indexing, meteorology and oceanography - the number of variables to measure can be unwieldy and at times even deceptive, because the underlying relationships can often be quite simple.

One such example is a spring-mass system shown in Figure 18. The engineer learned in school to model a system like this with an ideal massless spring, with all the mass at a point and no mass or damping in the spring. For an ideal system, when the ball is released a small distance away from equilibrium (i.e. the spring is stretched), the ball will oscillate along the length of the spring indefinitely at a set frequency.

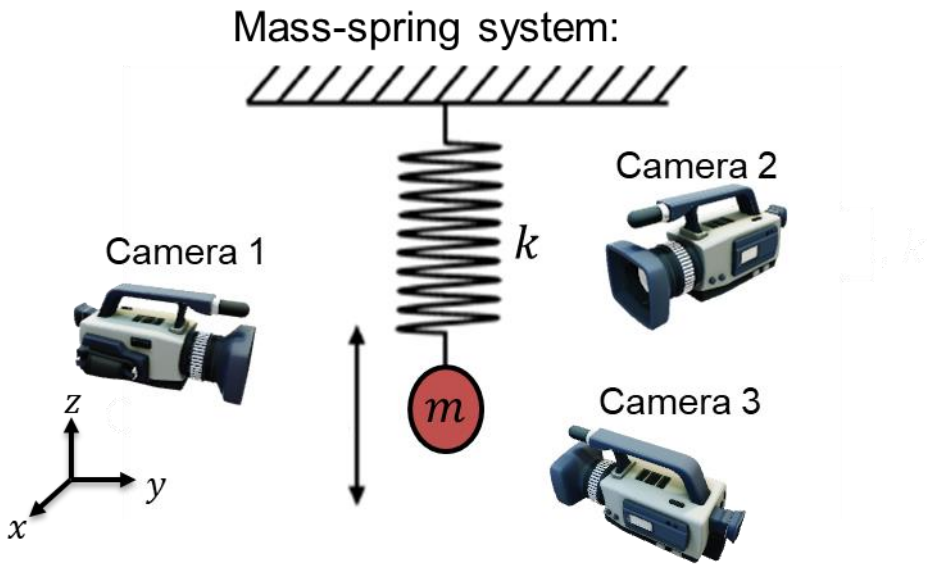


Figure 18 A spring-mass motion example. The position of a ball attached to a spring is recorded using three cameras 1, 2 and 3. The position of the ball tracked by each camera is depicted in each panel.

The actual system is not quite so ideal since the spring has some mass and there is some damping due to friction. The engineer decided to measure the ball's position in a three-dimensional space (since the world is inherently a three-dimensional world). He placed three video cameras around the spring-mass system and recorded the motion at 120 frames per second, which provided three distinct projections of the two-dimensional position of the ball. Unfortunately, the engineer placed the video cameras as three arbitrary locations, which meant that the angles between the measurements might not even be 90°! After recording for several minutes, the engineer was faced with the big question: how to get one-dimensional motion data from the two-dimensional projection data collected from three different locations?

This measurement and analysis challenge is often present. It is difficult to always know *a priori* which measurements best reflect the dynamics of the system, or else more data is recorded than actually needed.

One additional challenge is the noise and variability in the real-world system. In the spring-mass system being analyzed, air resistance, imperfect cameras positioning, and friction result in a less-than-ideal spring mass system. Noise contaminates our data set only serving to obfuscate the dynamics further. Keep this example in mind as we delve further into abstract concepts. In this textbook, we will use Singular value decomposition (SVD) and Principle component analysis (PCA) to solve this problem. The goal is to identify the most meaningful basis to re-express a data set. The hope is that this new basis will filter out the noise and reveal hidden structure. In the example of the spring, the explicit goal of PCA is to determine: "the dynamics are along the z-axis." In other words, the goal of PCA is to determine that z, i.e. the unit basis vector along the z-axis, is the important dimension. Determining this fact allows an experimenter to discern which dynamics are important, redundant or noise. After we identify the important dimension, we can estimate damping coefficient and effect of mass from the collected noisy data with the help of data science techniques such as genetic programming (see Figure 19).

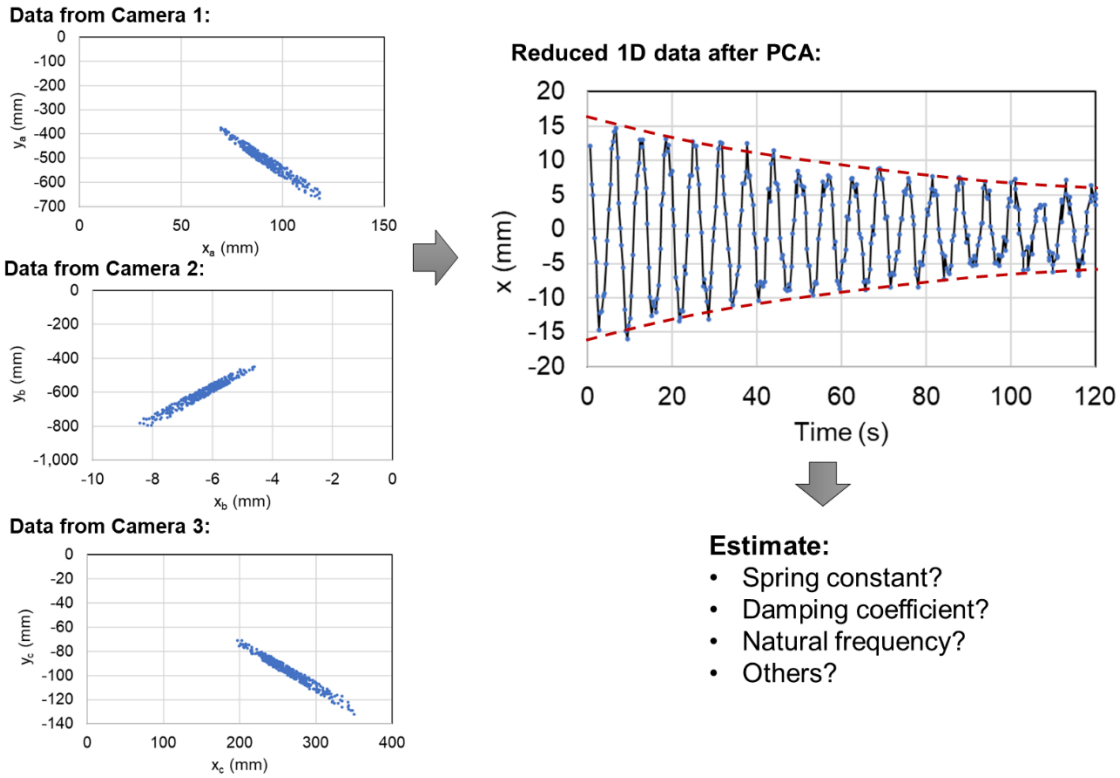


Figure 19 A example reducing high-dimensional data to 1D data and then estimate important parameters in the system.