Chapter 2
Multimodal Data Generation and Collection

Abstract Mechanistic data science is heavily reliant on the input data to guide the analysis involved. This data can come from many shapes, sizes, and formats. This process is a key part of the scientific process and generally involves observation and careful recording. Costly data collection from physical observation can be enhanced by taking advantage of the modern computer hardware and software to simulate the physical experiments and generate further complementary data. Efficient data collection and management through a database can expedite the problem- solving timeline and help in rapid decision-making aspects. This chapter shows data collection and generation from different sources and how they can be managed efficiently. Feature-based diamond pricing and material property testing by indentation are used to demonstrate key ideas.

Keywords Data collection · Data generation · Empiricism · Mechanism · Mechanistic · Database · Training · Testing · Cross-validation · High fidelity · Low fidelity · Multimodal · Multifidelity · Features · Macro-indentation · Micro-indentation · Nano-indentation · Microhardness · Hardness · Brinell · Vickers · Load-displacement · Sensing · Indenter

As the name suggests, *data* is the key input for mechanistic *data* science. The question though is "where does the data come from". The answer is "it can come from many sources and in many formats", which gives rise to the term *multimodal data collection and generation*.

As discussed in Chap. 1, scientific investigation starts with observation, which invariably leads to data collection to test hypotheses that are developed. Analysis of the data leads to a proven hypothesis and the discovery of new scientific theory. Collecting data from physical observation can be very costly and difficult to control independent variables, but it is possible to take advantage of the modern computer

Supplementary Information: The online version of this chapter (https://doi.org/10.1007/978-3-030-87832-0_2) contains supplementary material, which is available to authorized users.

W. K. Liu et al., *Mechanistic Data Science for STEM Education and Applications*,
https://doi.org/10.1007/978-3-030-87832-0_2

hardware and software to simulate the physical experiments and generate further complementary data. Efficient data collection and management through a database can expedite the problem-solving timeline and help in rapid decision making aspects. In this chapter, the data collection and generation process are discussed from different sources and how they can be managed efficiently. The data collection and generation process will be demonstrated through example problems such as diamond features and prices, and material property testing by indentation.

2.1 Data as the Central Piece for Science

Data provides evidence to supports the scientific knowledge and distinguish it from conjecture and opinion. As discussed in Chap. 1, this practice goes back centuries to the times of Copernicus, Kepler, Brahe, and Galileo. Galileo has been called the Father of the Scientific Method, in part for his structured use of data in his scientific pursuits. This can be illustrated in his classic beam problem. In his book *The Two New Sciences*, Galileo presented a drawing of a cantilever beam bending test as shown in Fig. 2.1.

Galileo's analysis centered on the question "*how forces are transmitted by structural members?*" To answer this question, his unique approach led him to a conclusion that holds true for all structural member used even today. His approach can be seen in the following four steps:

1. **Observation:** he observed that as the strength of the beam were affected by the length of the beam and the cross section of the beam.
2. **Hypothesis:** he noted that the beam strength decreased with length, unless the thickness and breadth were increased at an even greater rate.
3. **Testing and Data collection:** he performed many experiments on different size and shapes of the structural member and tried to collect data on their ability to carry and transmit loads.
4. **Scientific theory:** From the data and observations (understanding the mechanisms) he came to a conclusion which is applicable irrespective of the length, size, shape, materials for the structural member carrying loads. This also led to the scaling law that holds regardless of the size, shape, and material.

Galileo reported his finding as "the breaking force on a beam increases as the square of its lengths." A more familiar version of his findings is typically taught the undergraduate engineering students in a strength of materials class as the deflection formula for cantilever beam. The deflection of the tip of a beam can be related with the applied force (F), length of the beam (L), material property (Elastic modulus, E), and geometric factor (area moment of inertia for the beam cross-section, I). The equation of the tip deflection is $\delta = FL^3/3EI$, which works regardless of material, size, shape, and load.

Another example of data to empiricism or mechanism is the Kepler's three laws (1609–1619) of planetary motion. Kepler observed the solar system for many years,

116 THE TWO NEW SCIENCES OF GALILEO

has here been said the weight of the solid BD itself has been left out of consideration, or rather, the prism has been assumed to be devoid of weight. But if the weight of the prism is to be taken account of in conjunction with the weight E, we must add to the weight E one half that of the prism BD: so that if, for example, the latter weighs two pounds and the weight E is ten pounds we must treat the weight E as if it were eleven pounds.

Fig. 17

SIMP. Why not twelve?

SALV. The weight E, my dear Simplicio, hanging at the extreme end C acts upon the lever BC with its full moment of ten pounds: so also would the solid BD if suspended at the same point exert its full moment of two pounds; but, as you know, this solid is uniformly distributed through-

Second Day consists in a failure to see that, in such a beam, there must be equilibrium between the forces of tension and compression over any cross-section. The correct point of view seems first to have been found by E. Mariotte in 1680 and by A. Parent in 1713. Fortunately this error does not vitiate the conclusions of the subsequent propositions which deal only with proportions—not actual strength—of beams. Following K. Pearson (Todhunter's *History of Elasticity*) one might say that Galileo's mistake lay in supposing the fibres of the strained beam to be inextensible. Or, confessing the anachronism, one might say that the error consisted in taking the lowest fibre of the beam as the neutral axis. [*Trans.*]

Fig. 2.1 An excerpt from Galileo's *The Two New Sciences* [1]

based on his observations, came up with three laws to describe the motion of the planets in the solar system (see Fig. 2.2). Described succinctly, the laws are (1) the law of orbits, (2) the law of areas, and (3) the law of periods. All his laws are empirical in nature and describe the mechanism for planetary motion from direct observation of his collected data. This is the mechanistic part of this problem which explains the mechanisms of the planetary motions; however, these laws do not explain the reason behind such planetary motions. The science behind this is later discovered by Sir Isaac Newton through the law of gravity in 1687. The theory was further questioned by Einstein in his research from 1907 to 1917 in which he explained the motion of the planet Mercury and developed the theory of general relativity and gravity [2]. This remains the latest understanding of gravity and the motion of planets.

From these two examples of Kepler and Galileo, it can be seen that data comes from physical observation of the system and provides the basis for finding governing

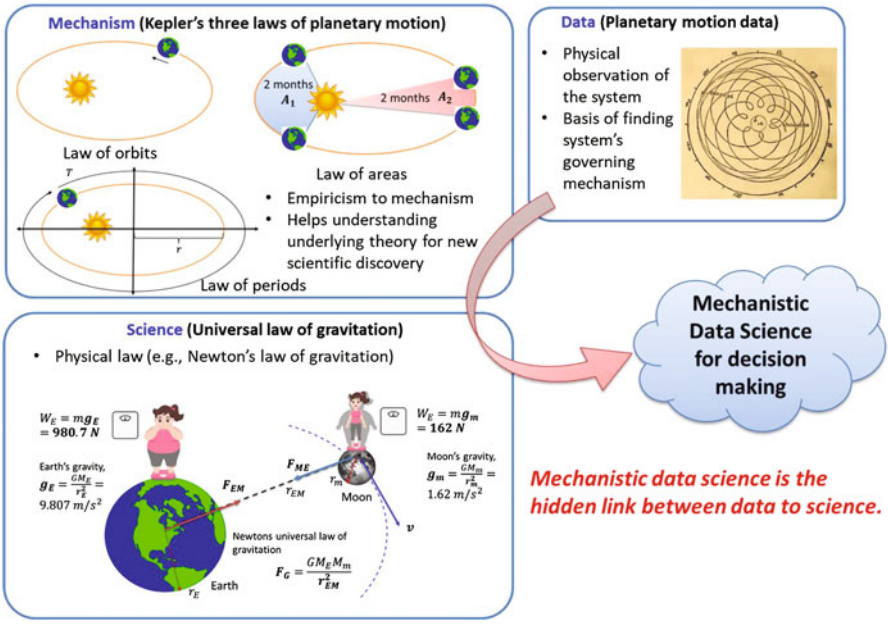

Fig. 2.2 Discovery of law of gravitation from planetary motion data. Gravity working among two different objects can be described by the Newton's universal law of gravitation. The gravity of earth and moon are 9.807 and 1.62 m/s^2, respectively. The weight is mass times the gravity force acting on her. If she measures her weight in moon, she will be definitely happy to see her weight loss. However, if she is intelligent enough, she will realize that it is her mass which matters not the weight

mechanisms. On the other hand, science explains the detailed reasoning behind such observations. The intermediate step is finding the mechanisms and justify the scientific hypothesis, which is the "Mechanistic" aspect of a problem. Combining data with the underlying scientific mechanism results in a unique scientific approach defined in this book as *Mechanistic Data Science*. The goal of mechanistic data science is twofold: (1) mining the data intelligently to extract the science, (2) combining data and mechanisms for decision making.

One can easily understand the amount of time (approximately 300 years from Kepler to Einstein) and effort necessary to develop science from just raw data observation with the devotion of great scientific minds. We can break down this process into two parts: Data to empiricism or mechanism and mechanism to science.

- **Data to empiricism or mechanism:** Collected data are analyzed and the relationship between data samples are established using mathematical tools and intuition.
- **Mechanism to science:** Once the mechanisms of a problem is clearly understood, the theory is further questioned to find the reasoning of such behavior found in nature.

Here mechanistic data science clearly establishes the links between the data and science through identifying the governing mechanisms. But it all starts from the data. In the next section, we will discuss about the data and some commonly used databases to search for data. However, finding appropriate data may be very challenging and sometimes consume years to find the appropriate data to solve the problem. Hence, having a clear idea on what data to collect and how to collect them make a significant difference on problem solving.

2.2 Data Formats and Sources

Data is a collection of information (numbers, words, measurements) or descriptions that describes a system or problem. It is an integral part of daily life, including financial data for tracking the stock market, climate data for predicting seasonal changes, or transportation data in the form of automobile accident records, train schedules, and flight delays. This information may take many forms – text, numbers, images, graphs, etc.—but it is all data.

Data is divided into two categories: qualitative and quantitative. *Qualitative* data is descriptive information. For example, saying "it's too hot outside" *describes* the temperature without giving an exact value. In contrast, *quantitative data* is numerical information. Saying "it's 90° outside" gives the precise temperature in terms of a numerical value but does not give context.

Quantitative data can be further divided into discrete or continuous data. *Discrete* data can only take certain values. For example, a dataset recording student heights has a fixed number of datapoints corresponding to one per student. If there are 10 students being measured, there must be 10 data points, not some fractional number like 10.7 data points. *Continuous* data can have any value within a given range. For example, temperature changes continuously throughout the day and can have any value (e.g. 47.783°, 65°, 32.6°). In summary, discrete data is counted, while continuous data is measured.

Data that is used for a mechanistic data science analysis can be obtained in multiple ways including measurements, computation or from existing databases.

- Measurement: this generally involves setting up a controlled experiment and instrumenting the test to measure data. A test can be repeated multiple times to evaluate consistency (e.g. does a coil of aluminum used to make soda cans meet the specifications) or can be conducted with a varying set of parameters (e.g. what is the effect of changing material suppliers). Making measurements has long been one of the key endeavors of science. For example, Chap. 1 gave a historical description of the data collection for planetary motion and falling objects and how that led to fundamental laws of science.
- Computation: in many cases there is a tremendous amount of mechanistic knowledge that can be used to compute data. For example, as described in the indentation example later in this chapter, finite element analysis can be used to

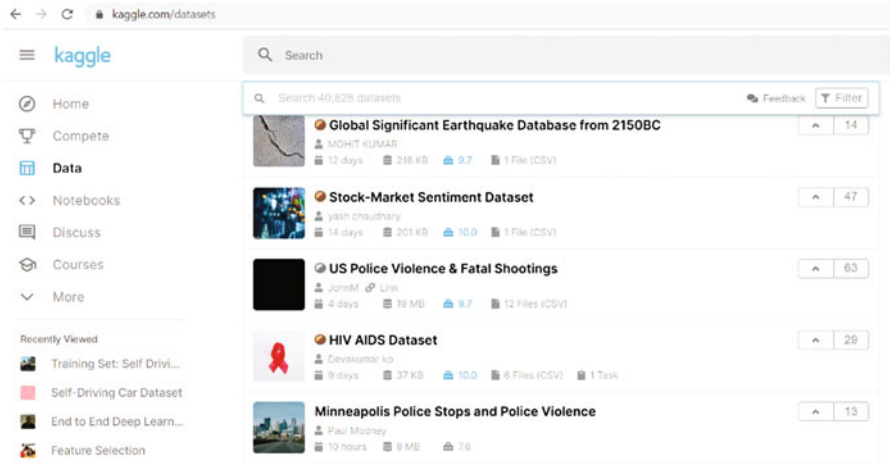

Fig. 2.3 Sample machine learning database Kaggle

perform detailed calculations of how a structure will perform under different loading conditions.

- Existing database: as data is collected, it can be compiled into a large dataset that can be used for reference or for further analysis.

A *database* is an organized collection of data, generally stored and accessed electronically from a computer system. Mechanistic data science relies on several engineering and machine learning databases, spanning a wide range of industries, disciplines, and problems. For example, **Kaggle** contains various datasets for machine learning, **Materials Project** contains materials data such as compounds and molecules, the **National Climate Data Center** (NCDC) contains datasets on weather, climate, and marine data, and the **National Institute of Standard and Technology** (NIST) has materials physical testing databases [3].

Figure 2.3 is a snapshot listing some Kaggle databases which can be used for data science and machine learning. As can be seen from this list, there is a wide range of data available, including the stock market, earthquakes, global diseases, and other engineering and social topics.

A typical dataset is composed of features and data. *Features* are distinctive variables that describe part of the data, and are typically arranged in columns, such as "country" or "earthquake magnitude" [4]. It should be noted that real data that is used for machine learning is not perfect; many "good" datasets are not complete, and often need to be prepared before being used for analysis. A dataset might often have noisy data, with some outliers that come from sensor errors or other artifacts during the data collection process. While it is very tempting to ignore or discard those outliers from the dataset, it is recommended that they be given careful attention before deciding how to treat them. It is the role of the data scientist to interpret those data and check the influence of those outliers on the hypothesis of the problem and population statistics of the data. Typically, regression-based models are well suited

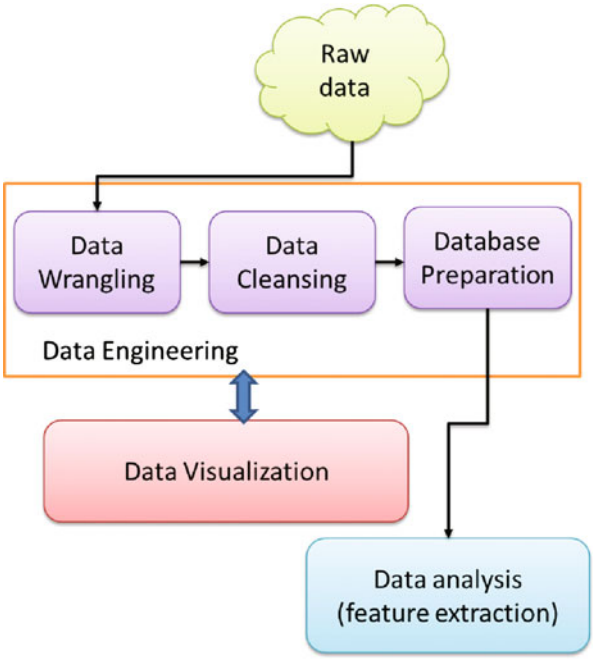

Fig. 2.4 Data preparation for analysis

to identify outliers in linearly correlated data; clustering methods and principal component analysis are recommended if the data does not show linearity on the correlation planes [5]. Regression models are discussed in Chap. 3; clustering techniques and principal component analysis are discussed in Chap. 5 of this book. Steps involved in data preparation are captured in Fig. 2.4 and described below [6]. The extraction of mechanistic features is discussed more extensively in Chap. 4.

- **Raw Data:** collected data in an unmodified form.
- **Data Wrangling:** transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for analysis.
  - Data wrangling prepares data for machine interpretation. For example, a computer may not recognize "Yes" and "No," so this data is converted to "1" and "0," respectively. The meaning of the raw data is unchanged, but the information is mapped to another form.
- **Data Formatting:** formatting data for consistency, associating text data with labels, etc.
- **Data Cleansing:** providing attributes to missing values and removing unwanted characters from the data.
- **Database Preparation:** adding data from more than one source to create a database.

While many of these steps are performed with automated data processing techniques, user input is still required. For example, after data is transformed, cleaned, and prepared, it should be visualized. Plotting data, displaying images, and creating graphs often reveal key trends, leading to better understanding of data. This knowledge allows data scientists to manually evaluate data science results, ensuring machine learning trends reflect data.

2.3 Data Science Datasets

Data science consists of using data to find a functional relationship between input and output data. As shown in Fig. 2.5, an input dataset $\boldsymbol{X_i^N}$ with i features is used for developing the functional relationship $\boldsymbol{y_j^N = f\left(X_i^N\right)}$.

The function development starts by dividing the input dataset $\boldsymbol{X_i^N}$ into a training set, a validation set, and a test set (Fig. 2.6), with the training set generally being the largest. The inputs and outputs from the *training set* are fit to a mapping function $\boldsymbol{f\left(X_i^N\right)}$ using regression analysis to develop a mechanistic data science model. The *validation set* measures the accuracy of the model after the training step. This process is repeated with the updated model until the error between the predicted output and the actual output is below a required threshold. Once the error is minimized, the final functional form is established, the function is evaluated against the *test set*. Choosing the training, testing, and validation set from the data can be done either randomly or systematically. One systematic approach uses K-fold cross validation, where the data set is divided in K number of bins and different bins are used for training, testing, and validation. Cross validation makes the model more

Fig. 2.5 Fitting dataset inputs and outputs to a functional form with machine learning

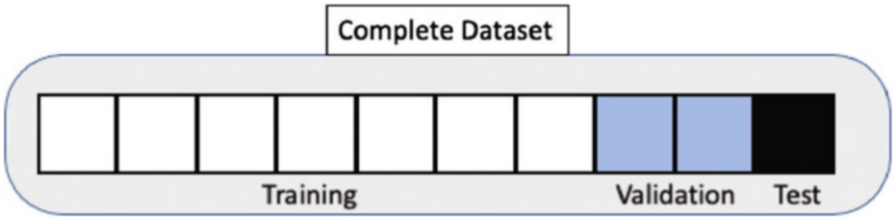

Fig. 2.6 Data division into training, validation, and test sets for machine learning. In this figure, the training set comprises 70% of the data, the validation set comprises 20% of the data, and the test set comprises 10% of the data

robust and removes bias in model training. More details on the *K*-fold cross validation are discussed in Chap. 3 and applied in the some of the examples in Chap. 7.

Data modality refers to the source (or mode) of the data. Mechanistic data science is able to incorporate *multimodal* data (data from multiple sources and test methods). Data deviations between various sources are resolved through calibration. For example, indentation data can be obtained through **physical experiments** and **computer simulations** [7]. An example of this is shown later in the chapter.

Data fidelity describes the degree to which a dataset reproduces the state and behavior of a real-world object, feature or condition. Fidelity is therefore a measure of the realism of a dataset [8]. It can be categorized as **high fidelity** or **low fidelity.** This is a somewhat subjective measure which depends on the application, but high fidelity data is generally more accurate and more expensive to obtain. For example, micro-indentation data is high fidelity compared to macro-indentation data, but low fidelity compared to nanoindentation data.

Machine learning techniques can improve the resolution of low fidelity data, transforming it into high fidelity data without the large collection cost [9].

2.4 Example: Diamond Data for Feature-Based Pricing

Diamond pricing analysis using regression techniques is shown in Chap. 1. Diamonds can be described with several features such as cut, color, clarity, and carat, and the price of a diamond is a function of all of these features. A dataset was downloaded from Kaggle containing data for **53,940 diamonds** with **10 features**. A sample of this dataset is shown in Fig. 2.7. For this diamond dataset to be used for predicting prices based on features, the input feature index, $i = 9$, represents the number of independent variables (in this example, features including cut, color, clarity, and carat). Similarly, the output feature index, $j = 1$, represents the number of dependent variables (i.e. price). The number of data points in the dataset is $N = 53{,}940$.

Cut, color, clarity, and carat are four features known as the 4 C's. They are defined as:

- Cut: the proportions of the diamond and the arrangement of surfaces and facets.
- Color: color of the diamond, with less color given a higher rating
- Clarity: the amount of inclusions in a diamond
- Carat: the weight of the diamond

Some diamond features such as cut, color, and clarity are not rated using numerical values. They must be converted to numerical values in order to be used in a calculation. In this case, the cut, color, and clarity are assigned numerical values based on the number of individual classifications for each.

carat	cut	color	clarity	depth	table	x	y	z	price
0.23	Ideal	E	SI2	61.5	55	3.95	3.98	2.43	326
0.21	Premium	E	SI1	59.8	61	3.89	3.84	2.31	326
0.23	Good	E	VS1	56.9	65	4.05	4.07	2.31	327
0.29	Premium	I	VS2	62.4	58	4.2	4.23	2.63	334
0.31	Good	J	SI2	63.3	58	4.34	4.35	2.75	335
0.24	Very Good	J	VVS2	62.8	57	3.94	3.96	2.48	336
0.24	Very Good	I	VVS1	62.3	57	3.95	3.98	2.47	336
0.26	Very Good	H	SI1	61.9	55	4.07	4.11	2.53	337
0.22	Fair	E	VS2	65.1	61	3.87	3.78	2.49	337
0.23	Very Good	H	VS1	59.4	61	4	4.05	2.39	338
0.3	Good	J	SI1	64	55	4.25	4.28	2.73	339
0.23	Ideal	J	VS1	62.8	56	3.93	3.9	2.46	340
0.22	Premium	F	SI1	60.4	61	3.88	3.84	2.33	342
0.31	Ideal	J	SI2	62.2	54	4.35	4.37	2.71	344
0.2	Premium	E	SI2	60.2	62	3.79	3.75	2.27	345
0.32	Premium	E	I1	60.9	58	4.38	4.42	2.68	345
0.3	Ideal	I	SI2	62	54	4.31	4.34	2.68	348
0.3	Good	J	SI1	63.4	54	4.23	4.29	2.7	351
0.3	Good	J	SI1	63.8	56	4.23	4.26	2.71	351
0.3	Very Good	J	SI1	62.7	59	4.21	4.27	2.66	351
0.3	Good	I	SI2	63.3	56	4.26	4.3	2.71	351

Fig. 2.7 A sample of data extracted from diamond features and prices

Cut Rating	Numerical value
Premium	1
Ideal	2
Very Good	3
Good	4
Fair	5

Clarity Rating	Numerical value
IF—Internally Flawless	1
VVS1,2—Very, Very Slightly Included 1,2	2
VS1,2—Very Slightly Included 1,2	3
SI1,2—Slightly Included 1,2	4
I1—Included 1	5

The color rating scale ranges from D to Z, where D is colorless and Z is a light yellow or brown color. For the given dataset, the diamond colors ranged from D to J and the numerical values were as assigned as: Color (D, E, F, G, H, I, J) → (1, 2, 3, 4, 5, 6, 7).

Once all the feature data is converted to numerical values, data normalization can be performed to scale all the data features from 0 to 1 if a regression analysis is to be performed (this will be discussed in more detail in Chap. 4).

2.5 Example: Data Collection from Indentation Testing

Material hardness testing by indentation is a multimodal data collection technique. Hardness testing consists of pressing a hardened tip into the surface of a material with a specified load and measuring the dimensions of the small indentation that is made. The indentation is generally very small and, as such, the test is considered non-destructive. Furthermore, since the resistance to surface indentation is related to the stress required to permanently deform the material, the measured hardness can often be correlated to other material properties like the ultimate tensile strength.

Indentation testing varies with sample size and shape, but the fundamental process remains the same. As shown in Fig. 2.8, the indenter is pressed into the surface of the material with a specified force and leaves a small impression. The hardness is determined by measuring the size of the indent for the applied load [10].

Macro-indentation is used to test large samples, with applied load exceeding 1 kgf. Small samples are tested using *micro-indentation*, using applied load ranging from 1–1000 gf. For even smaller scales, *nanoindentation* (also known as *instrumented indentation*) is used. For the nanoindentation scale, the applied load is less than 1 gf [11]. Common indenter tips (see Fig. 2.9) include hemispherical balls (used for the Brinell hardness test) and various pointed tips (used for the Vickers hardness test and nanoindentation test).

The load vs. indentation depth for a typical nanoindentation test is plotted in Fig. 2.10 [13]. The decrease in the indentation depth when the load is removed is determined by the *elasticity* of the material. The sample results in Fig. 2.10 show some elasticity since the final (or residual) depth, h_r, is less than maximum depth, h_m. The net result is that the indenter tip leaves a permanent impression of depth h_r in the surface of the material due to localized surface deformation [12].

The contact area of the indentation depends on the indentation depth and indenter shape. Figure 2.11 shows several indenter tips and the corresponding contact area equations, where d is the indentation depth [12].

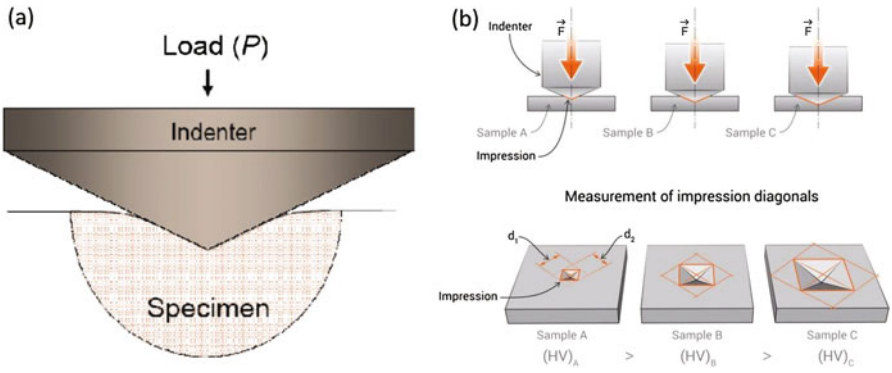

Fig. 2.8 (**a**) Indentation testing experimental set-up and (**b**) impression data (https://matmatch.com/learn/property/vickers-hardness-test)

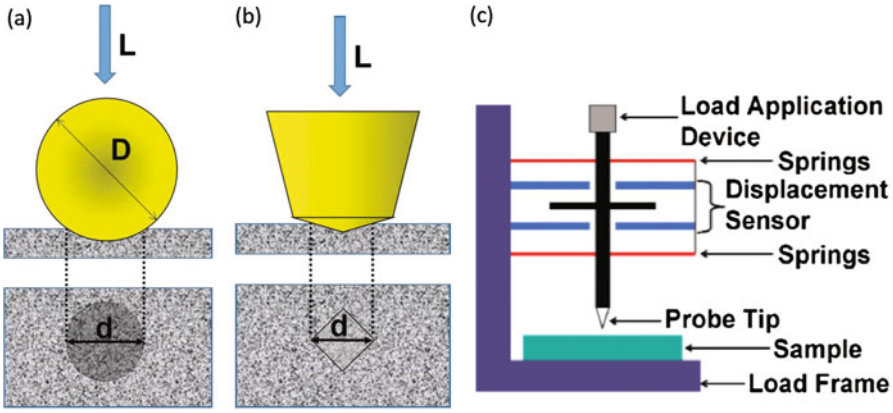

Fig. 2.9 Experimental set-ups for (**a**) Brinell hardness test, (**b**) Vickers microhardness test, and (c) nanoindentation test [11, 12]

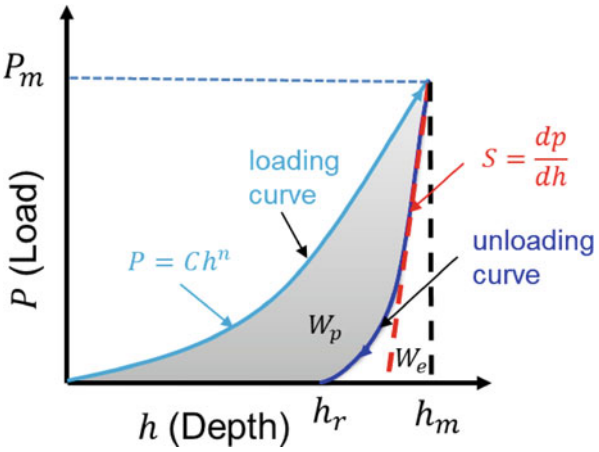

Fig. 2.10 Load vs. indentation depth data generated by nanoindentation testing

The indentation data can come from experiments and computer simulations (see Fig. 2.12). The experimental data is collected using imaging and sensing.

- *Experimental* data is obtained through the indentation. Typically, indentation experiments record the load-displacement data.
  - High resolution Atomic Force Microscopes (AFM) are used for *imaging* of the indented surface. The surface fracture pattern provides critical information on the material deformation during the indentation process. Additionally, the contact area of the indenter can be measured from these high-resolution microscope images.

Parameter	Berkovich	Cube-corner	Cone	Spherical	Vickers
Shape					
C-f angle	65.35°	35.264°	—	—	68°
Projected Contact area	$24.5600d^2$	$2.5981d^2$	πa^2	πa^2	$24.5044d^2$

Fig. 2.11 Common indenter tips and corresponding contact area equations [12]

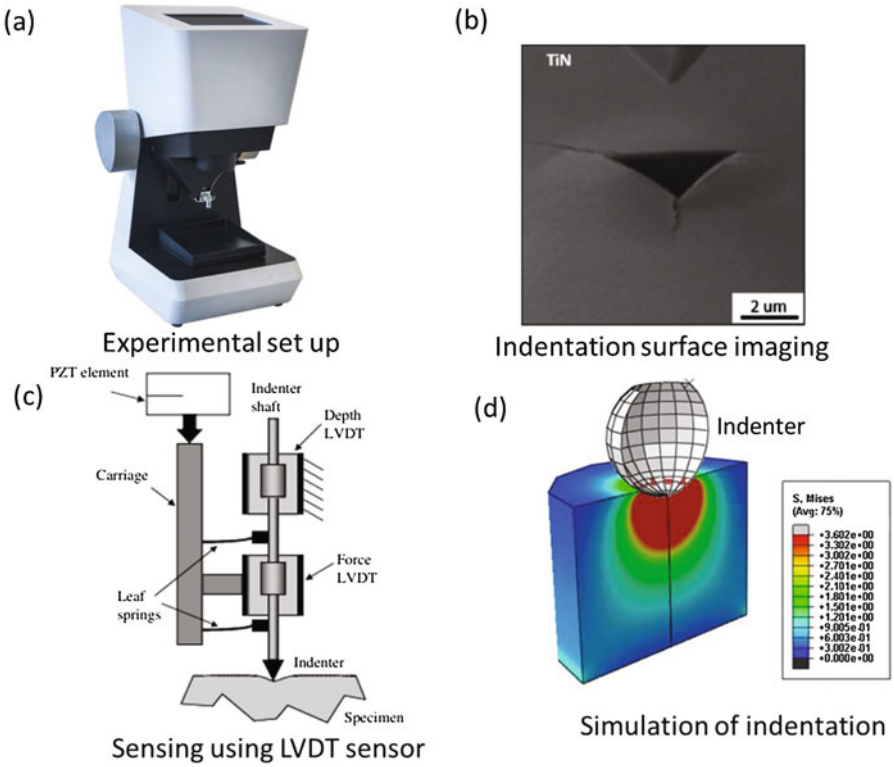

Fig. 2.12 Indentation data sources: (**a**) experiments, (**b**) imaging, (**c**) sensing using LVDT sensor, (**d**) computer simulation using FEM (a video is available in the E-book, Supplementary Video 2.1) [14]

– *Sensing* is accomplished using various force and displacement measurements. A common displacement sensor is the Linear Variable Differential Transformer (LVDT), which measures the movement of the indenter shaft through

electric voltage change and provides the load and displacement data. Other displacement measurement techniques include differential capacitors or optical sensors. Force can be measured through a spring-based force actuation system.

- *Computer simulations* are powerful tools to compute the load displacement data and materials behavior. Computational simulation methods, such as the Finite Element Method (FEM), has been used extensively to compute mechanical properties of materials through indentation simulation. FEM is a well-known computer simulation method for computing deformation and stress given the geometry and the material properties. It has successfully replaced or augmented physical testing for many areas of engineering product development.

Computer simulations of surface indentation can also provide valuable data for material characterization. A physical test result and a finite element computer simulation are shown in Fig. 2.13. With proper calibration, the two methods produce nearly identical triangular indentations in the material and the simulation can be used to provide additional insight and data for the indentation process.

Machine learning databases often combine information for different modes of data collection and levels of fidelity. For example, Table 2.1 summarizes

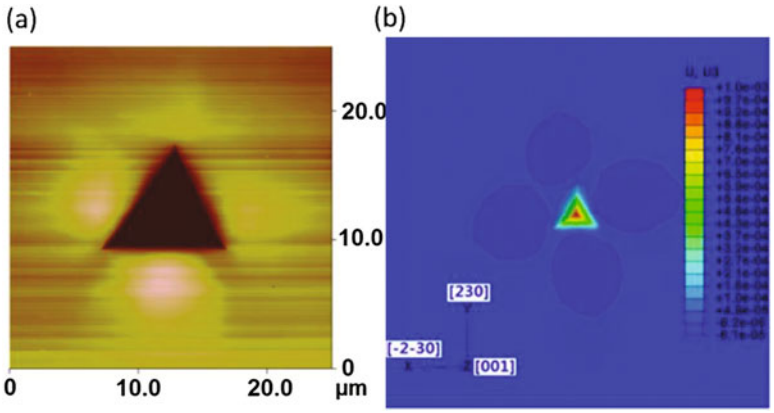

Fig. 2.13 Indentations produced by (**a**) physical nanoindentation experiment, and (**b**) finite element method computer simulation [7]

Table 2.1 Summary of nanoindentation testing data [15]

Material	Experiment	Computation
Al-6061 alloy	7 experiments	**2D FEM (Axisymmetric):** 100 simulations each for conical indenter half angle of 50, 60, 70, 80° **3D FEM:** 15 simulations for Berkovich indenter
Al-7075 alloy	7 experiments	
3D printed Ti-6Al-4V alloys (six samples)	144 experiments for each sample	Not available

nanoindentation testing data for three different materials [15]. The aluminum alloy data consisted of 422 load-displacement curves (7 physical tests, 400 2D axisymmetric FEM simulations, and 15 3D FEM simulations). The data for the 3D printed Ti-6Al-4V material consisted of 864 load-displacement curves (144 experiments on each of six samples).

The mix of experimental and computational data represent different *modalities (or sources)*. In addition, the 2D and 3D FEM simulations also represent different levels of *fidelity (or resolution)*. The 3D FEM simulations are more comprehensive but are computationally intensive. The 2D axisymmetric simulations assume that the indentation is axisymmetric but afford a much higher level of model refinement. Consequently, the fidelity of 2D and 3D simulation must be understood within the context of the physical test being modeled.

2.6 Summary of Multimodal Data Generation and Collection

Mechanistic data science analysis frequently utilizes multimodal and multi-fidelity data as one data source rarely provides sufficient data to fully represent an engineering problem. Experimental data obtained through direct observation is considered the most reliable but may be too expensive or complicated to obtain. Limited experimental data may need to be supplemented with simulations or published experimental results. As a result, data scientists must identify, collect, and synthesize required information from a variety of modes and fidelities to solve engineering problems. This idea will be further developed in Chap. 3 Optimization and Regression.

References

1. Galileo (1638) The two new sciences
2. Siegfried T (2015) Getting a grip on gravity: Einstein's genius reconstructed science's perception of the cosmos. Science News
3. Badr W (2019) Top sources for machine learning datasets. Towards Data Science. https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b. Accessed 1 Sep 2020
4. Kumar M (2020) Global significant earthquake database from 2150BC. Kaggle [Online]. https://www.kaggle.com/mohitkr05/global-significant-earthquake-database-from-2150bc. Accessed 23 June 2020
5. Salgado CM, Azevedo C, Proença H, Vieira SM (2016) Noise versus outliers. Secondary analysis of electronic health records, pp 163–183
6. Jones MT (2018) Data, structure, and the data science pipeline. IBM Developer. [Online]. https://developer.ibm.com/articles/ba-intro-data-science-1/. Accessed 1 Sep 2020

7. Liu M, Lu C, Tieu K et al (2015) A combined experimental-numerical approach for determining mechanical properties of aluminum subjects to nanoindentation. Sci Rep 5:15072. https://doi.org/10.1038/srep15072
8. SISO-REF-002-1999 (1999) Fidelity Implementation Study Group Report. Simulation Interoperability Standards Organization. Retrieved January 2, 2015
9. Lu L et al (2020) Extraction of mechanical properties of materials through deep learning from instrumented indentation. Proc Natl Acad Sci 117(13):7052–7062
10. Wikipedia (2020) Indentation hardness. [Online]. Available: https://en.wikipedia.org/wiki/Indentation_hardness. Accessed 8 Sep 2020
11. Broitman E (2017) Indentation hardness measurements at macro-, micro-, and nanoscale: a critical overview. Tribol Lett 65(1):23
12. VanLandingham MR (2003) Review of instrumented indentation. J Res Natl Inst Stand Technol 108(4):249–265
13. Nanoindentation. Nanoscience instruments [Online]. https://www.nanoscience.com/techniques/nanoindentation/. Accessed 1 Sep 2020
14. Rzepiejewska-Malyska KA, Mook WM, Parlinska-Wojtan M, Hejduk J, Michler J (2009) In situ scanning electron microscopy indentation studies on multilayer nitride films: methodology and deformation mechanisms. J Mater Res 24(3):1208–1221
15. Extraction of mechanical properties of materials through deep learning from instrumented indentation. GitHub. [Online]. https://github.com/lululxvi/deep-learning-for-indentation. Accessed 23 Jun 2020