Contents

- What is data?
- Machine learning databases
- Terminology
  - Data preparation, Data modality, Data fidelity
- Data formats and sources
  - Experiments, Imaging, Sensing, Modeling and simulation
- Examples
  - Diamond data for feature-based pricing
  - Data collection from indentation testing

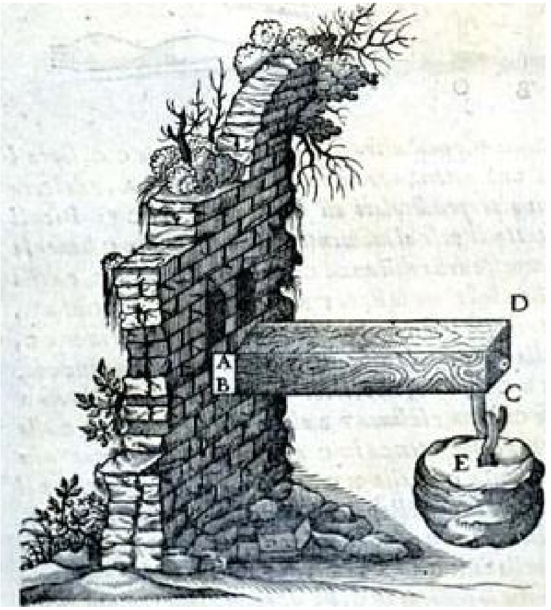
Data

- Key input for mechanistic data science
- Where does the data come from? It can come from many sources and in many formats. → multimodal data collection and generation
  - Physical observation: very costly and difficult to control independent variables
  - Modern computer HW and SW: simulate the physical experiments and generate further complimentary data
- Efficient data collection and management through a database
  - Expedite the problem-solving timeline → Help in rapid decision making

- Goal of mechanistic data science (MDS)
  - Mining the data intelligently to extract the science
  - Combining data and mechanisms for decision making

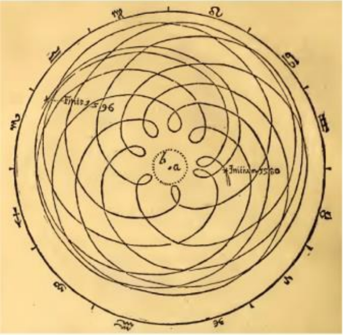
Data is the central piece for science

- Question: how are forces transmitted by structural members?
- Galileo's approach:
  - Data collection: performed many experiments on how size and shape of structural members affects their ability to carry and transmit loads
  - Observations : as length of a beam increases, its strength decreases, unless you increase the thickness and breadth at an even greater rate
  - Science: This led Galileo to recognize what we now call the scaling problem, there are limits to how big nature can will break under their own weight.
  - $deflction$: $\delta = \frac{FL^3}{3EI}$ This formula to calculate deflection of cantilever beams works for macroscopic beams made with all materials, size, shape and loads



Data is the central piece for science

- Evolution of scientific discovery: from data to empiricism to mechanism
  - Astronomical data: Observations of planetary orbits



Kepler's three laws of planetary motion (1609-1619)
- The Law of Orbits
- The Law of Areas
- The Law of Periods

- Newton's Laws of Motion (1687)

$$\begin{cases} F = ma \\ F = \dfrac{GM_1M_2}{r^2} \end{cases}$$

Data

- Physical observation of the system
- Basis of finding system's governing mechanism

Mechanism

- Empiricism to mechanism
- Helps understanding underlying theory for new scientific discovery

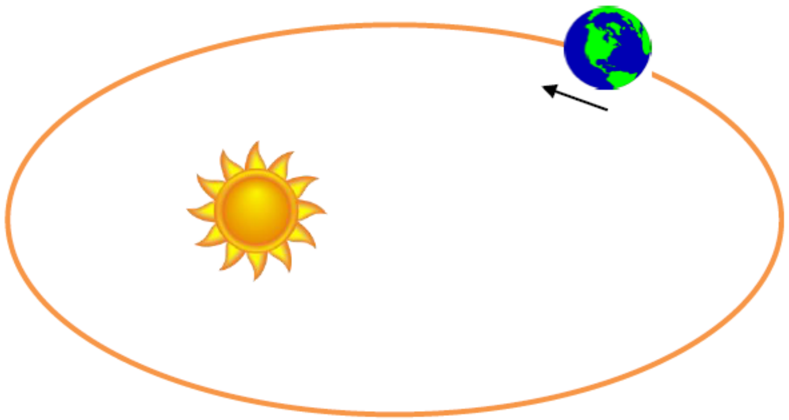
Science

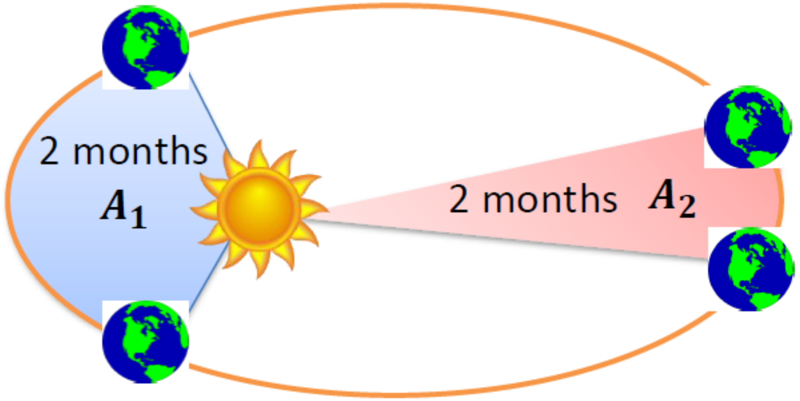
- Physical law (e.g., Newton's law of gravitation)

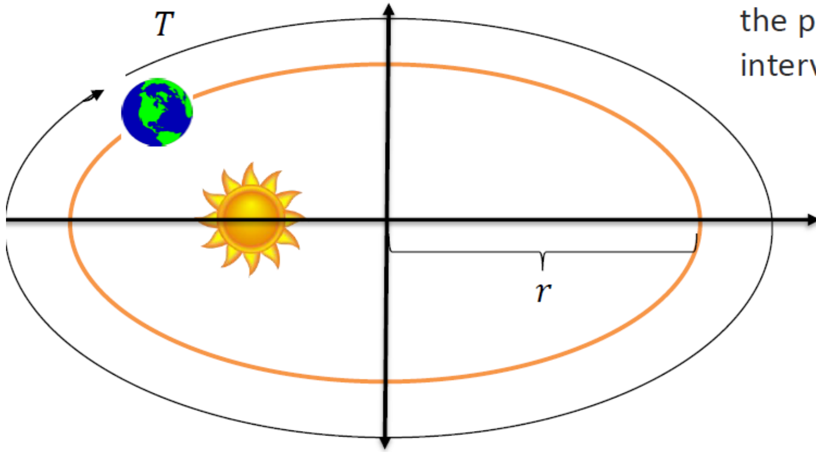
- Mechanistic data science is the hidden link between data to science

Kepler's Law (Data to Mechanism)

- First law (law of orbits): Each planet revolves around the sun in an elliptical orbit with the sun situated at one of the two foci.
- Second law (law of areas): The real velocity of a planet around the sun remains constant, OR, The radius vector drawn from the sun to the planet sweeps out equal areas in equal intervals of time.
- Third law (law of periods): The square of the time period period($\boldsymbol{T}$)of revolution of a planet around the sun is proportional to the cube of the semi major axis ($\boldsymbol{r}$)of its elliptical orbit.

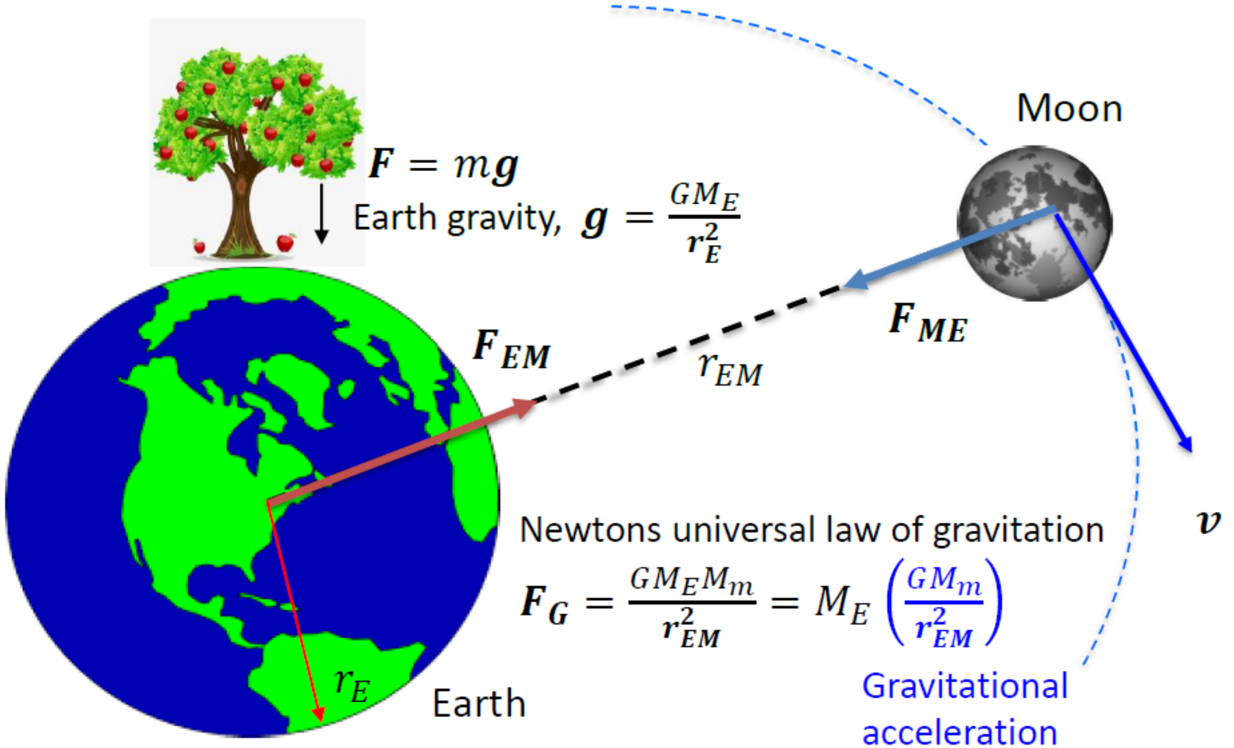






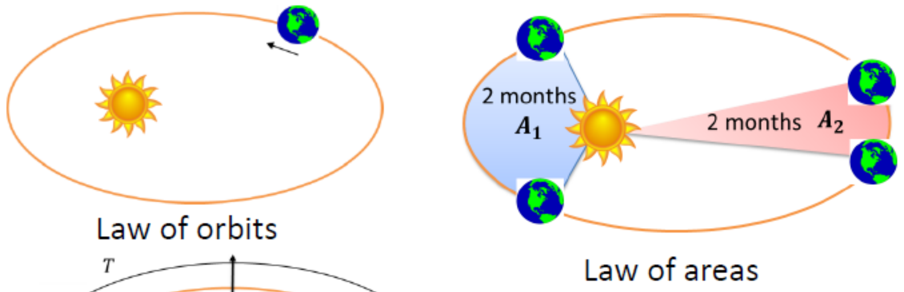
Mechanism to Science: Discovery of gravity

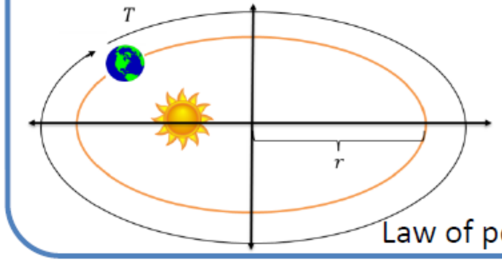
- Newtons universal law of gravitation
  - Every point mass attracts every single other point mass by a force acting along the line intersecting both points. The force is proportional to the product of the two masses and inversely proportional to the square of the distance between them.
- Force on a falling object (apple from a tree) in earth due to gravity is given by $\boldsymbol{F}=m\boldsymbol{g}$



Discovery of gravitation from planetary motion data

Mechanism (Kepler's three laws of planetary motion)

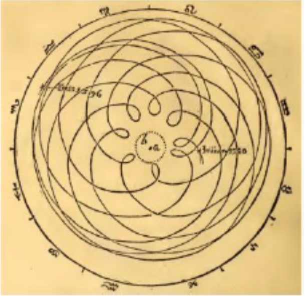




- Empiricism to mechanism
- Helps understanding underlying theory for new scientific discovery

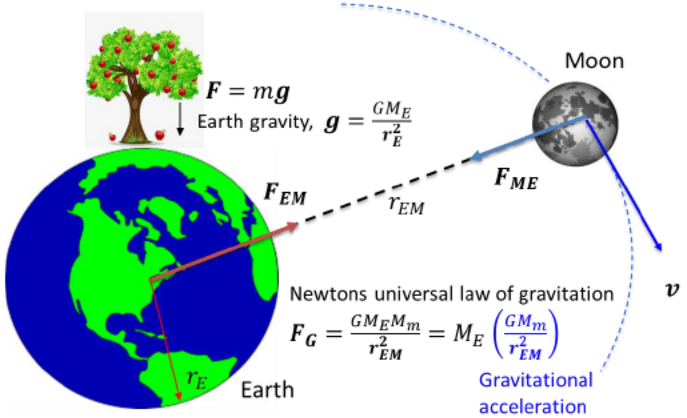
Data (Planetary motion data)

- Physical observation of the system
- Basis of finding system's governing mechanism



Science (Universal law of gravitation)

- Physical law (e.g., Newton's law of gravitation)



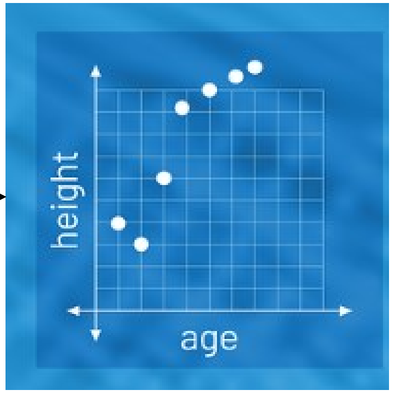
Mechanistic Data Science for decision making

Mechanistic data science is the hidden link between data to science.

What is Data?

- Data: collection of information (numbers, words, measurements, observations) or descriptions of things

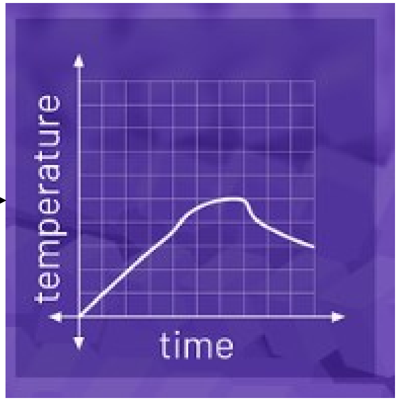
Data
- Qualitative
  "It's too hot outside"
  Descriptive
- Quantitative
  "It's 90°F outside"
  Numerical



Counted data

Discrete data can only take certain values (ex. only whole numbers)

7 data points = 7 people. You can't measure height for "7.3" people.



Measured data

Continuous data can take any value (within a range)

Temperature can take any value within Earth's range. It changes continuously, forming an infinite curve.

- Text, numbers, images, graphs, and signals are all common forms of data
- Data represents all industries and problems: finance, climate, transportation, etc

https://www.mathsisfun.com/data/data.html

Common Databases for Machine Learning Applications

- Database: organized collection of data, generally stored and accessed electronically from a computer system

Kaggle (www.kaggle.com)
- Machine learning datasets
- Open source
- Anyone can upload data
- Wide range of topics

NIST (National Institute of Standard & Technology) (https://materialsdata.nist.gov/)
- Materials physical testing database

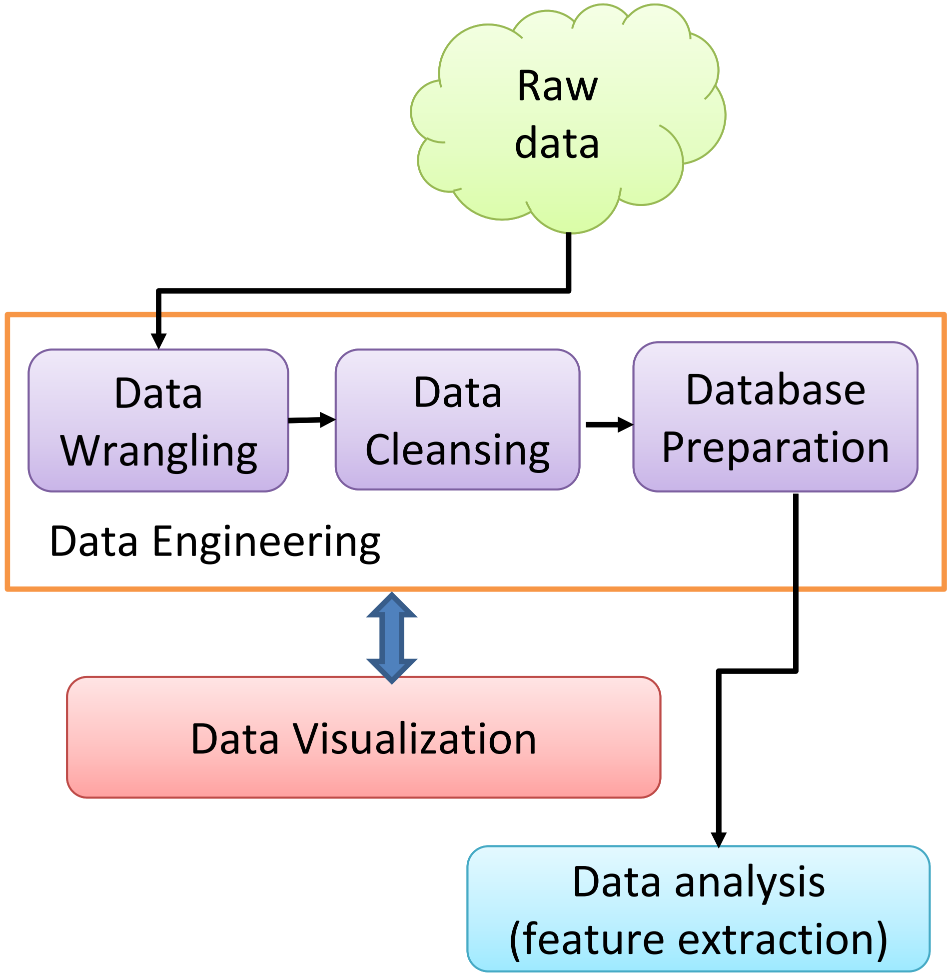
NCDC (National Climate Data Center) (https://www.ncdc.noaa.gov/cdo-web/)
- Weather and climate database
- Daily weather data
- Local climate data
- Marine data

Materials Project (https://materialsproject.org/)
- Materials database
- Materials data for : 144,595 inorganic compounds
- 63,876 molecules
- 530243 nanoporous materials

https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b

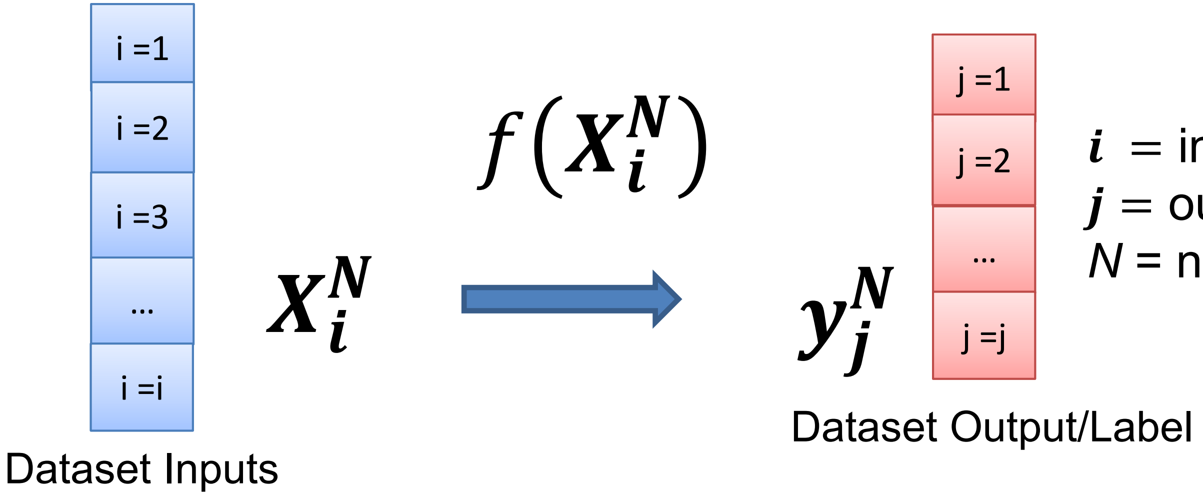
Data Preparation for Analysis

- **Raw data**: collected from the source directly
- **Data wrangling**: mapping and transforming raw data to another format for machine interpretation (ex. map Yes/No to 1/0)
- **Data Formatting**: formatting data for consistency (ex. formatting text data with labels)
- **Data Cleaning**: providing attributes to missing values and removing unwanted characters from the data
- **Database preparation**: adding data from 1+ sources to build your own database
- **Feature Extraction**
  - Identification of important features in the data
  - Determined with human expertise



https://developer.ibm.com/articles/ba-intro-data-science-1

Dataset for Machine Learning (1)



$\boldsymbol{i}$ = input feature index
$\boldsymbol{j}$ = output feature index
N = number of data points/samples

- $f(\boldsymbol{X}_i^N)$ maps input $\boldsymbol{X}_i^N$ to output $\boldsymbol{y}_j^N$.
- Machine Learning Goal = find the functional form $\boldsymbol{y}_j^N = f(\boldsymbol{X}_i^N)$

- For Kaggle diamond dataset:
  - i=1…9 (Carat, Cut, etc.)
  - j=1 (Price)
  - N=53,940 (number of diamonds sampled)

Dataset for Machine Learning (2)

$$\boldsymbol{X_i^N} \xrightarrow{f\left(\boldsymbol{X_i^N}\right)} \boldsymbol{y_j^N}$$

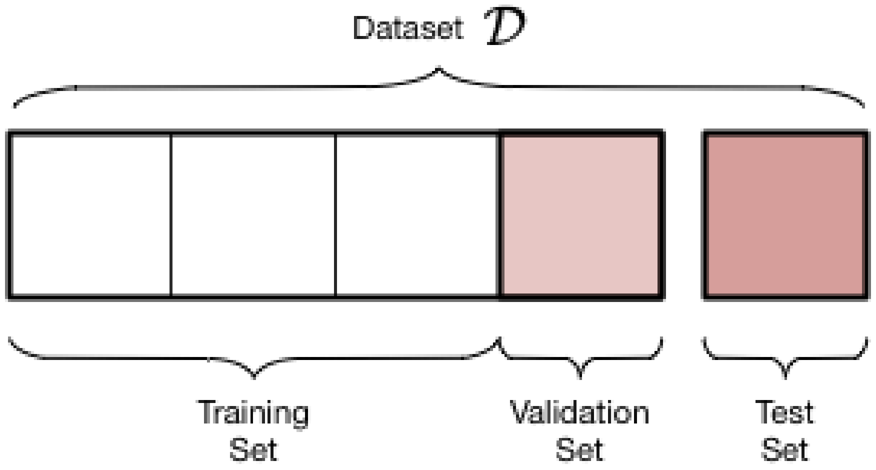
Dataset Inputs — Dataset Output/Label

$\boldsymbol{i}$ = input feature index

$\boldsymbol{j}$ = output feature index

N = number of data points/samples

- The dataset is divided into training (70%), validation (15%), and testing (15%) sets to find the functional relationship and confirm it is the **best possible fit**
- This process is **iterative**. The model is repeatedly trained, validated, and trained
- Final performance on the testing set is evaluated when function error is minimal



70% 15% 15%

- ❑ Training set: Inputs and outputs fit to mapping function $f\left(\boldsymbol{X_i^N}\right)$
- ❑ Validation set: Evaluate function (frequently after each training step)
- ❑ Testing set: Evaluate the final function $f\left(\boldsymbol{X_i^N}\right)$

Example: How can we identify a high quality diamond at a reasonable price?

1. The Pink Star

Image Source: Cosmopolitan Italia

Price: $71 million
Sold: April 2017 at Sotheby's Auction
Carat Weight: 59.6 carats
Color: Pink

2. Oppenheimer Blue Diamond

Image Source: Christie's

Price: $57.5 million
Sold: May 2016
Carat Weight: 14.62 carats
Color: Blue

3. Graff Vivid Pink Diamond

Image Source: Diamondhistorygirl

Price: $46 million
Sold: November 2010
Carat Weight: 24.78 carats
Color: Pink

4. The Princie Pink Diamond

Image Source: DailyMail.co.uk

Price: $39.3 million
Sold: April 2013
Carat Weight: 36.45 carats
Color: Pink

5. The Orange

Image Source: NY Post

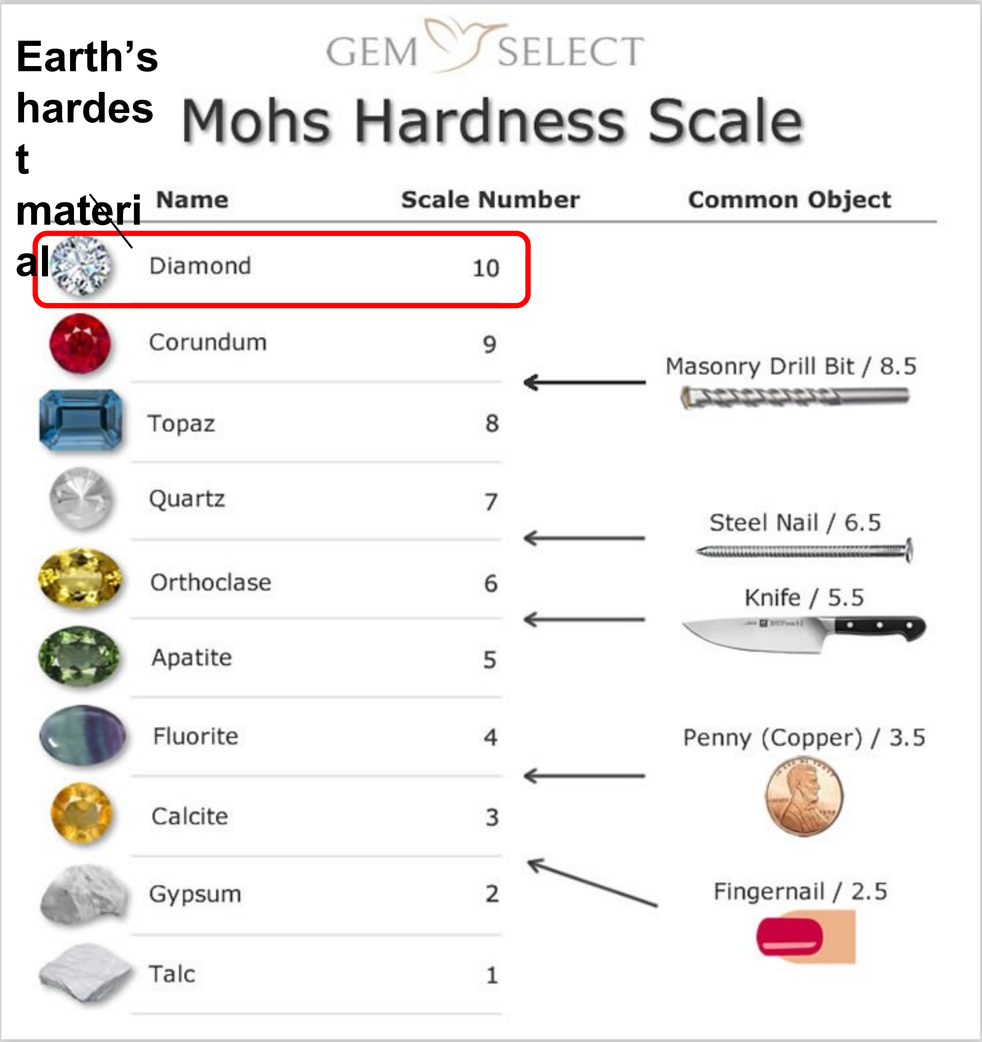
Price: $35.54 million
Sold: November 2013
Carat Weight: 14.82 carats
Color: Orange

6. The Largest Diamond Ever Sold

Image Source: CNBC

Price: $30.6 million
Sold: Christie's in 2013
Carat Weight: 118.28 carat
Color: Colorless

Mohs Scale of Hardness



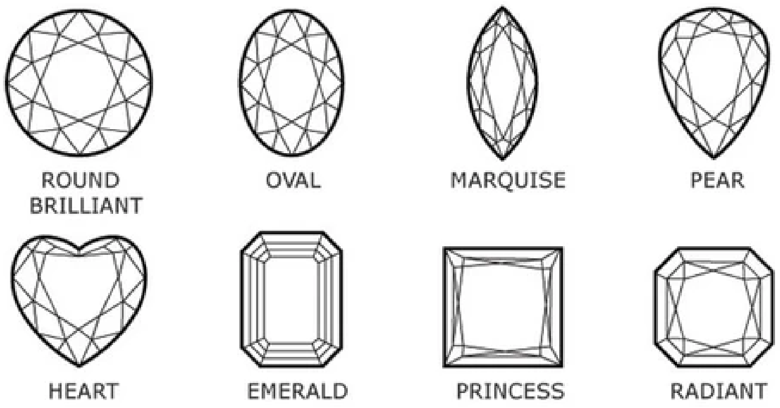
https://www.gemselect.com/gem-info/gem-hardness-info.php

- German mineralogist Frederick Mohs (1773-1839)
- How to Perform the MOHS Test?
  - ***Scratch it!***

https://geology.com/minerals/mohs-hardness-scale.shtml

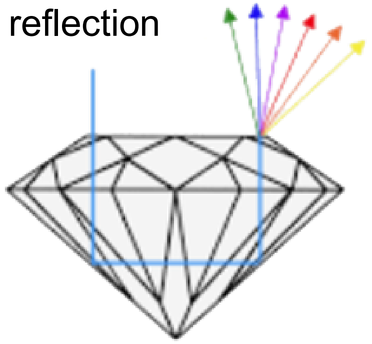
Features used to Characterize Diamonds

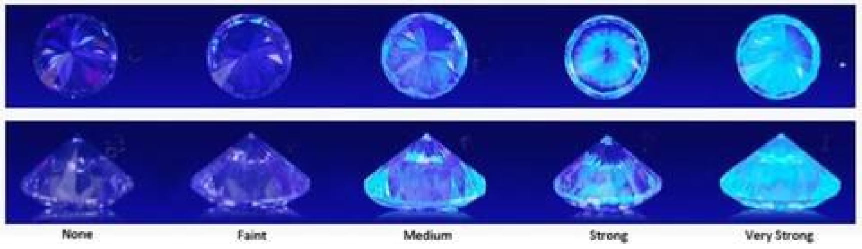
1. SHAPE
2. SIZE (carats)
3. CLARITY
4. COLOR
5. CUT
6. BRIGHTNESS
7. FIRE (dispersion)
8. SPARKLE
9. POLISH
10. SYMMETRY
11. FLUORESCENCE
12. DURABILITY
13. LUSTER



"One carat“ (100 points) equals the weight of 1/5 of a gram

Total internal reflection





Data Science focuses on quantifiable features

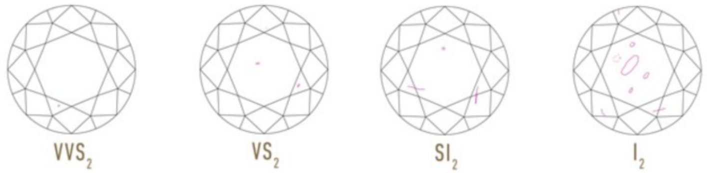
https://perrywinkles.com/blogs/education/introduction-to-diamonds

Different features can represent the same problem

4C Standard

COLOR

CLARITY



CUT

Excellent, Very good, Good, Fair, Poor

CARAT WEIGHT

"One carat“ (100 points) equals the weight of 1/5 of a gram

Gemological Institute of America (GIA)

Example: Diamond Data for Feature-based Pricing

- Kaggle (Datasets/Diamonds)
  - 53,940 diamonds with 10 features

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39

Price: ($326--$18,823)
Carat: (0.2--5.01)
Cut: (Fair, Good, Very Good, Premium, Ideal)
Color: (J (worst) to D (best))
Clarity: (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
Size in x direction in mm (0--10.74)
Size in y direction in mm (0--58.9)
Size in z direction in mm (0--31.8)
Depth: **z** / mean(x, y) = 2 * **z** / (x + y) (43--79) (%)
Table: width of top of diamond relative to widest point (43--95) (%)

Color (D, E, F, G, H, I, J) → (1, 2, 3, 4, 5, 6, 7)
D: colorless ~ Z: light yellow or brown

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

Example: Moneyball

- Kaggle (Datasets/Moneyball)
  - MLB statistics 1962-2012
  - Billy Beane and Paul DePodesta, Oakland Athletics, 2002
  - 1,232 data with 15 features
    - Player: Batting average (BA), runs batted in (RBI)
    - Win 95 games to make the playoffs, score 133 more runs than opponents
    - On-base percentage (OBP), slugging percentage (SLG)=(1B+2B*2+3B*3+HR*4)/AB, on-base plus slugging (OPS)=OBP+SLG

Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
ARI	NL	2012	734	688	81	0.328	0.418	0.259	0			162	0.317	0.415
ATL	NL	2012	700	600	94	0.32	0.389	0.247	1	4	5	162	0.306	0.378
BAL	AL	2012	712	705	93	0.311	0.417	0.247	1	5	4	162	0.315	0.403
BOS	AL	2012	734	806	69	0.315	0.415	0.26	0			162	0.331	0.428
CHC	NL	2012	613	759	61	0.302	0.378	0.24	0			162	0.335	0.424
CHW	AL	2012	748	676	85	0.318	0.422	0.255	0			162	0.319	0.405

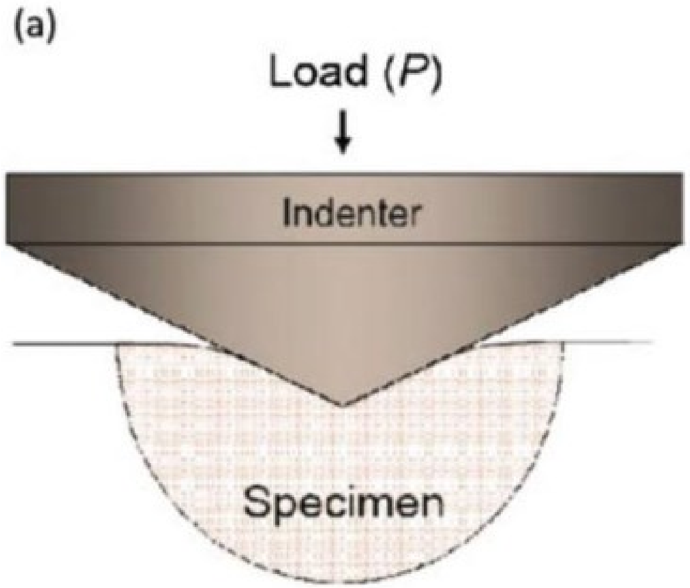
Example: Data Collection from Indentation Testing

Mechanistic Data Science
Applications: Materials Engineering

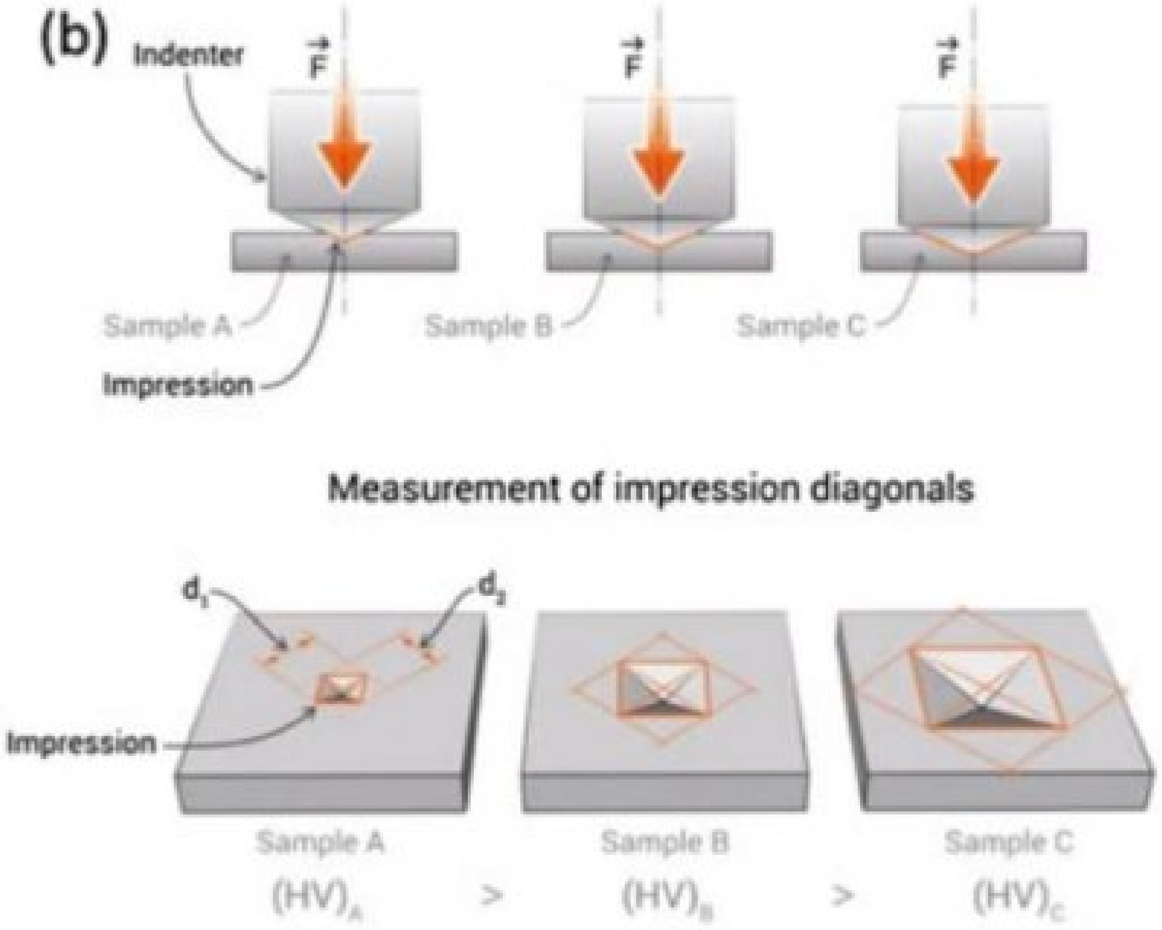
→ **How to analyze material properties?**

- Indentation testing: material testing for hardness
- Hardness: resistance to penetration of a hard indenter (related to material strength)
- Significance
  - Testing is simple, fast, relatively inexpensive, and not destructive
  - Hardness is closely related to critical mechanical properties: strength, ductility, and fatigue resistance
  - More plausible at small scales then tensile test

Example: Data Collection from Indentation Testing



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	$24.5044d^2$

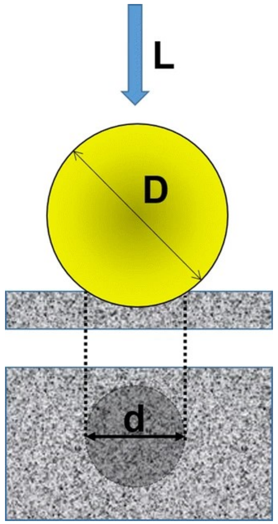


https://www.astm.org/Standards/E140.htm

https://en.wikipedia.org/wiki/Indentation_hardness

Indentation Tests: Vary by Sample Size and Shape

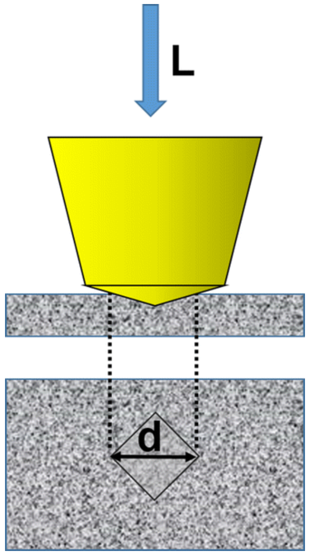
Macroindentation



Brinell *macrohardness* test

- Applied load > 1kgf
- Example: Vickers, Brinell, Knoop, Janka, Meyer, Rockwell, Shore hardness test

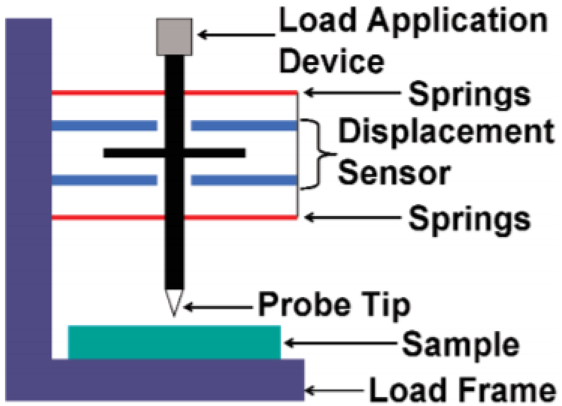
Microindentation



Vicker *microhardness* test

- Applied load = 1~1000 gf
- Example: Vicker, Knoop, microhardness test

Nanoindentation



Nanoindentation test

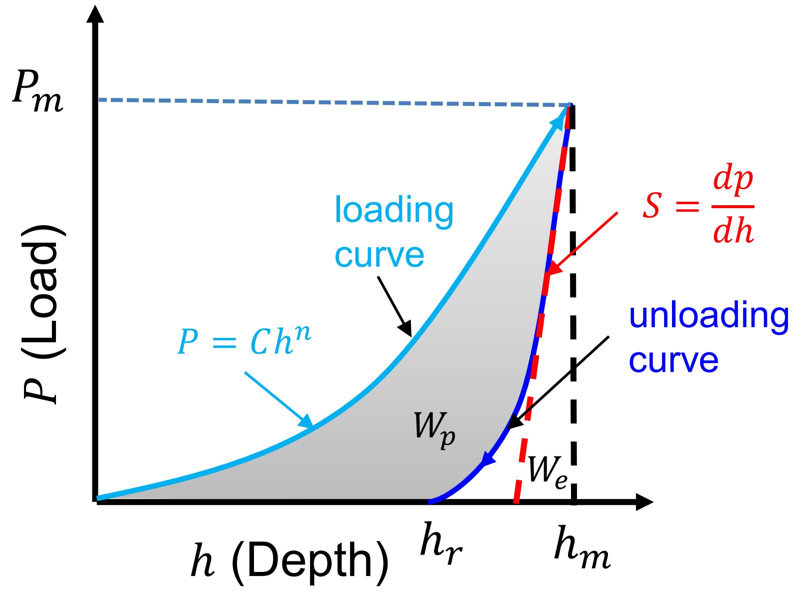
- Also known as instrumented indentation
- Applied load < 1gf
- Material testing at micro and nanoscale

Mark, J. Res. Natl. Inst. Stand. Technol. 108, 249-265 (2003)
Broitman, E., 2017, *Tribology Letters*, *65* (1), 23

Calculating Hardness from Load-Displacement Data

- What data we collect from indentation test?
- Hardness is calculated with the maximum applied load and the indenter contact area

P = applied load
h = indentation depth
S = slope of unloading curve
C = curvature of load-displacement curve
P_m = maximum load
h_m = maximum indentation depth
h_c = critical indentation depth (difficult to measure and must be calculated)
A = contact area (depends on indenter shape)
H = hardness (GPa)
E = elastic modulus (GPa)
ε = constant (depends on indenter geometry)
W_e= Elastic work done
W_p= Plastic work done



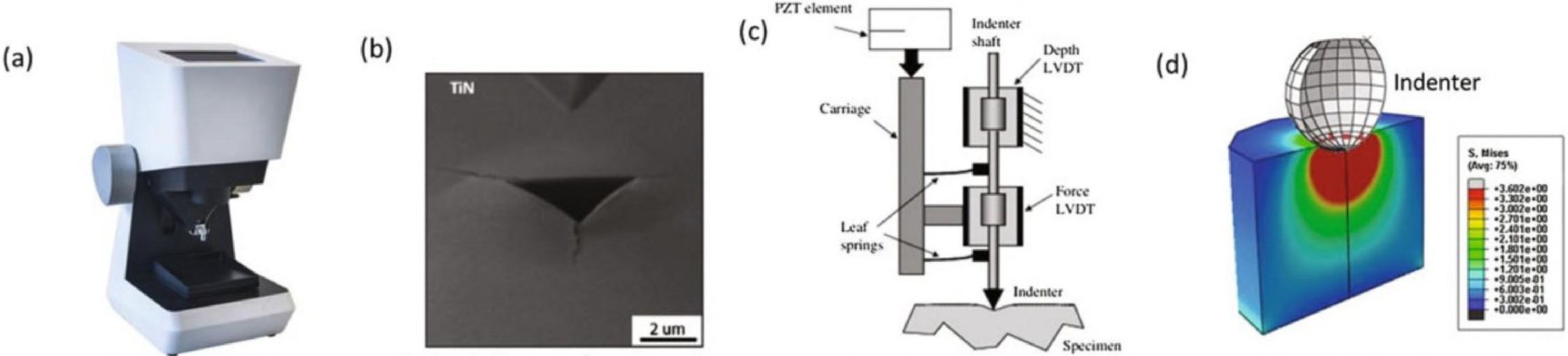
Key Equations

- $h_c = h - \frac{\varepsilon P_m}{S}$
- $A = f(h_c) = 24.56h^2$
- $\boldsymbol{H} = \frac{\boldsymbol{P_m}}{\boldsymbol{A}}$
- $E^* = \frac{S}{2h\beta}\sqrt{\frac{\pi}{24.56}}$ ($\beta = 1.034$ for Berkovich indenter)

https://www.nanoscience.com/techniques/nanoindentation

Indentation Data from Different Sources

- Data modality: data from different sources (modes)
- Data sources
  - **Experiment**: Instrumented indentation
  - **Imaging** : Scanning Electron Microscope (SEM)
  - **Sensors**: Load and displacement sensing using Linear variable differential transformer (LVDT)
  - **Modeling and simulation**: Finite element, Atomistic simulation

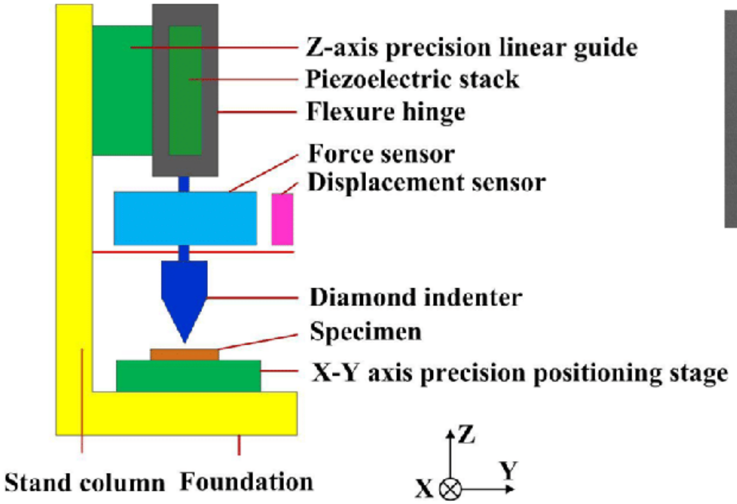


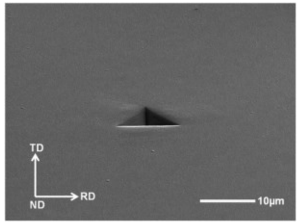
Experimental set up | Indentation surface imaging | Sensing using LVDT sensor | Simulation of indentation

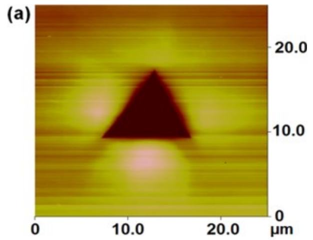
Indentation Data from Different Sources

- Load-displacement data can be found through **physical experiments** and **computer simulations**
  - Both experiment and simulations produce the same indentation
  - Deviations in modality require data calibration

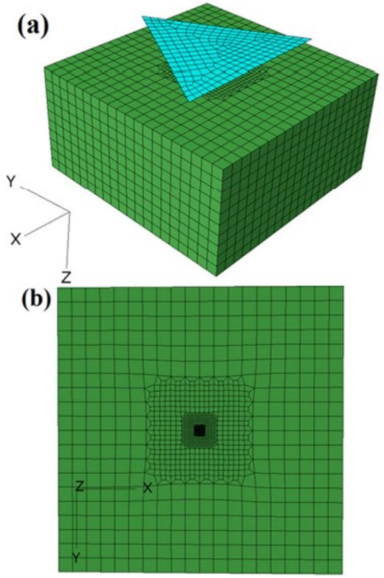
Nanoindentation Experiment and Imaging

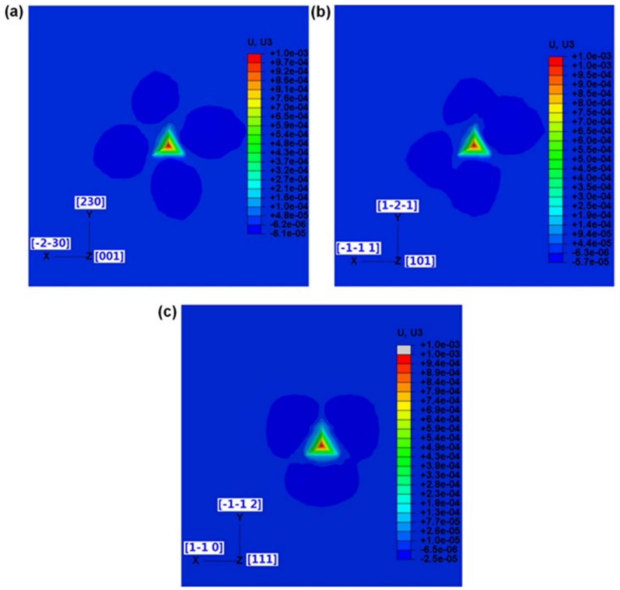






Finite Element Method / Computer Simulation



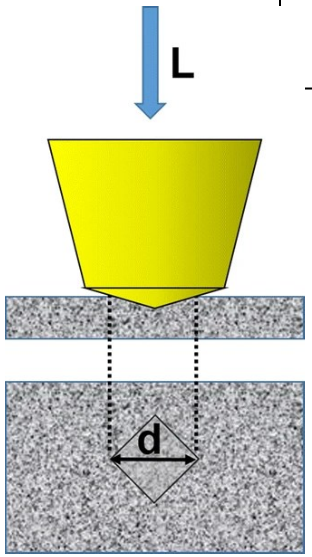


https://www.nature.com/articles/srep15072

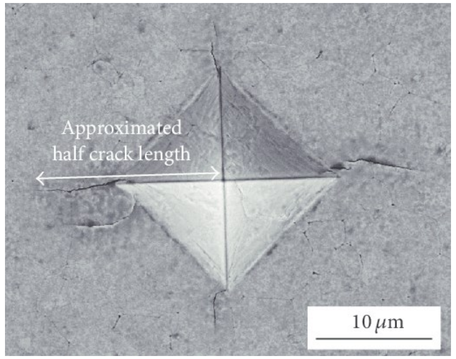
Indentation Data at Different Scales

- Data Fidelity: Resolution of data
  - High fidelity data is more accurate, but expensive
  - Machine learning improves the accuracy of low fidelity data, translating it to high fidelity with less cost
  - High and low fidelity are relative

Low-fidelity = Large Scale:
Microindentation



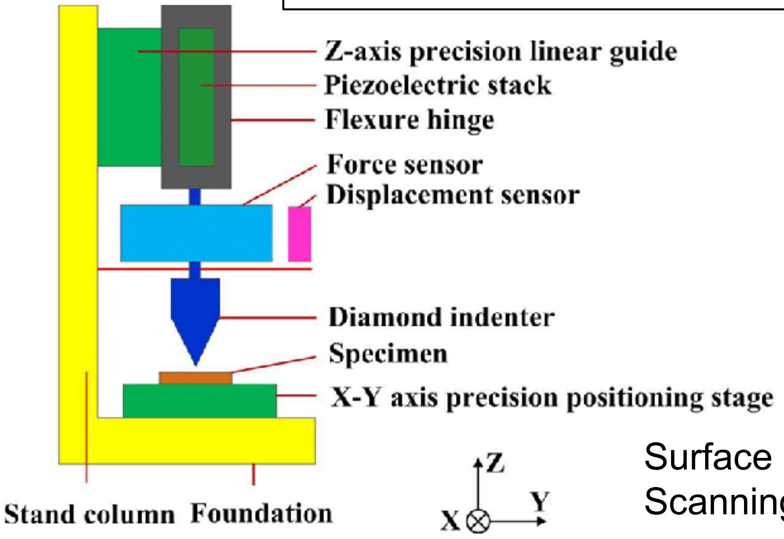
Vickers microindentation test



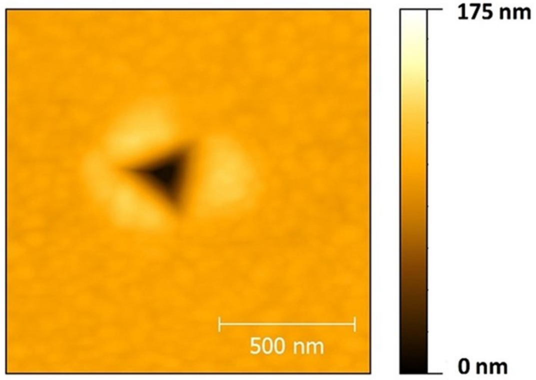
Impression shows indenter shape

Micron Resolution

High-fidelity = Small Scale:
AFM Nanoindentation



Nano Resolution



Surface imprint after nanoindentation,
Scanning electron microscope (SEM) image

Phan, Thien Q., et al. *Journal of Metallurgy* 2016 (2016)

https://www.nature.com/articles/srep17164

Indentation Database

Nanoindentation database summary:

Data Modality: Experimental and simulation data collected for each material

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

*Load-displacement curves are available for each experiment and simulation.

Data Fidelity: 2D (low fidelity) and 3D (high fidelity) simulations

Input data

- Load-displacement curvature
- Indentation depth
- Indenter shape and size
- Maximum load
- Unloading curve slope

→

Output data

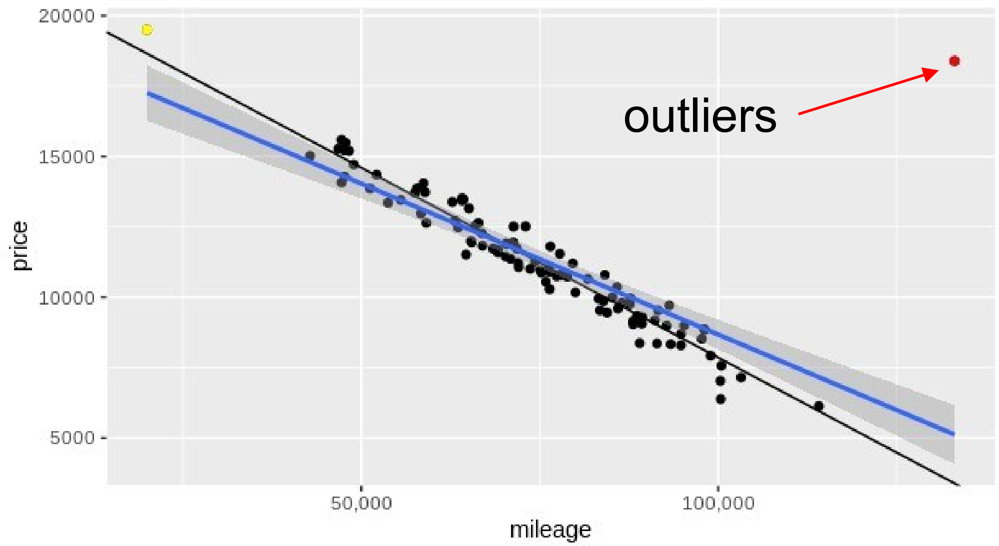
- Hardness
- Elastic modulus
- Yield strength

Lu, L., et al., Proceedings of the National Academy of Sciences, 2020, 117(13), 7052-7062.

Database source: https://github.com/lululxvi/deep-learning-for-indentation

Working with Noisy Data and Outliers

- Noise in the data is very common
- Source of noise: Human error in measurement, sensor fluctuation and so on
- Outliers are data coming from same source but vary significantly from other measurement



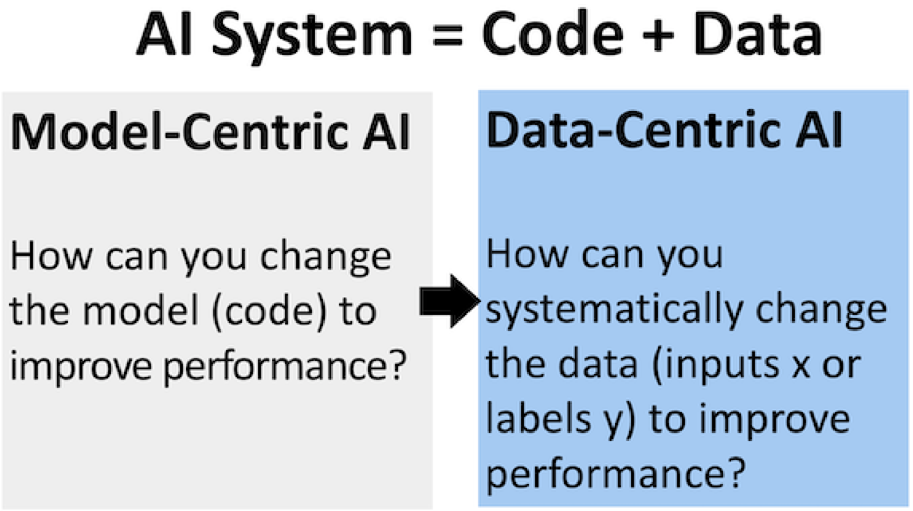
How do we know if this outlier to ignore or not?

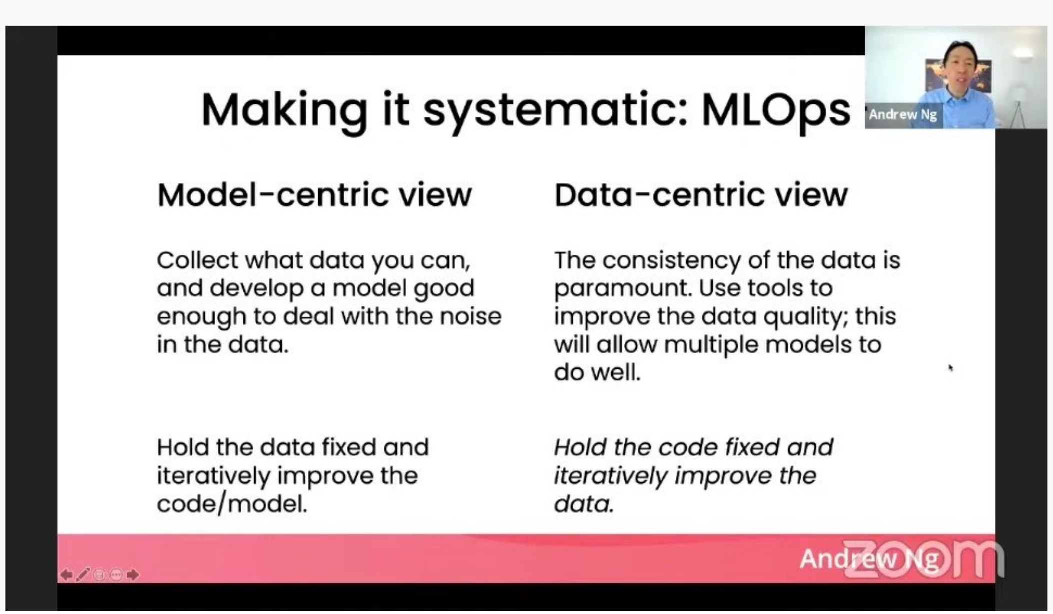
- Regression model can give idea on the data trend and help identify outliers or noise from the data.
- Type of regression:
  - Least square method
  - Lasso regression
  - Ridge regression
  - Elastic net regression, etc.

Challenges: Data

- Data on demand
  - Do not have enough data to run an ML model
  - Produce the data by using physics-based simulations
  - Issue: extensible or adaptive sampling is critical
- Data in hand
  - Use of the historical data, stored in different places and formats created by different software versions
  - Issue: reliability, need to be converted to metadata
    - Non-standard formats without proper access (op2, d3plot, bdf, …)
    - Non-uniform data (shell, solid, mesh, time series, text, …)
    - Inconsistent data (1d, 2d and 3d mixed)
    - Highly dirty data (oscillations, instability, …)
- Data in flight
  - Internet of Things (IoT) sensors: large amounts of fast data from operation
  - Issue: volume and quality of data

From Model-centric to Data-centric AI





https://www.youtube.com/watch?v=06-AZXmwHjo

Homework #1: Data Plotting

- Diamond
  - Fig.1.12 price vs. (a) carat, (b) separated by cut
- Moneyball
  - Fig.3.10 (a) RS vs. BA, OBP, SLG, OPS, (b) W vs. BA, OBP, SLG, OPS

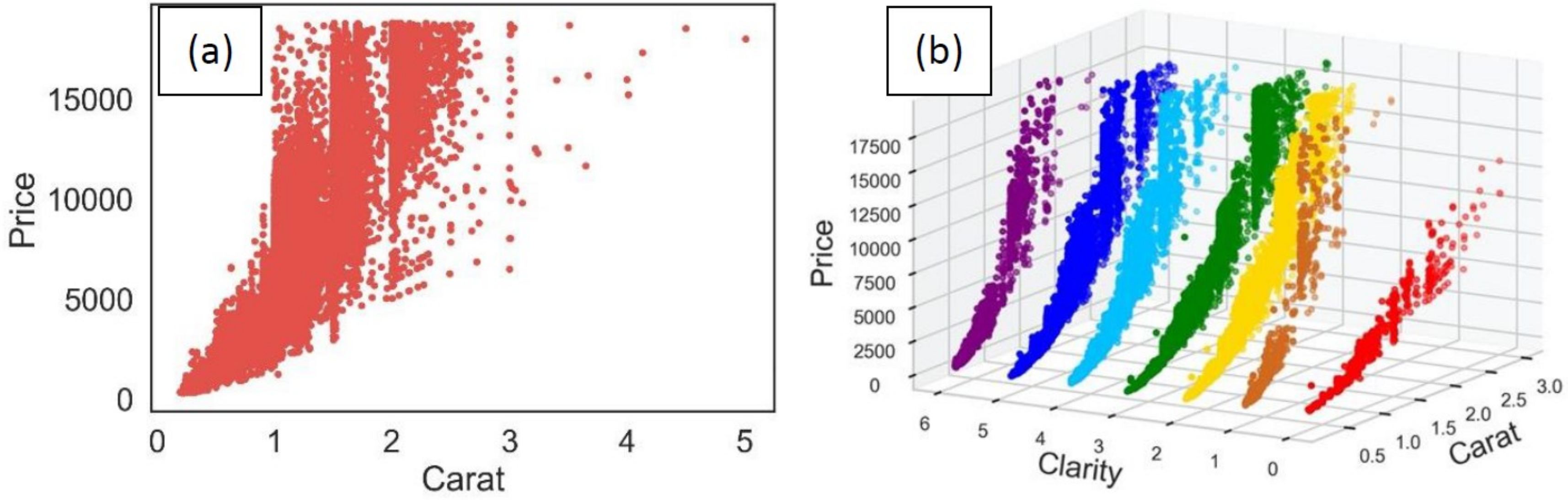


Figure 12 Diamond price vs carat (a) parameters combined (b) separated by clarity.