Week 6: Probability plots and data smoothing

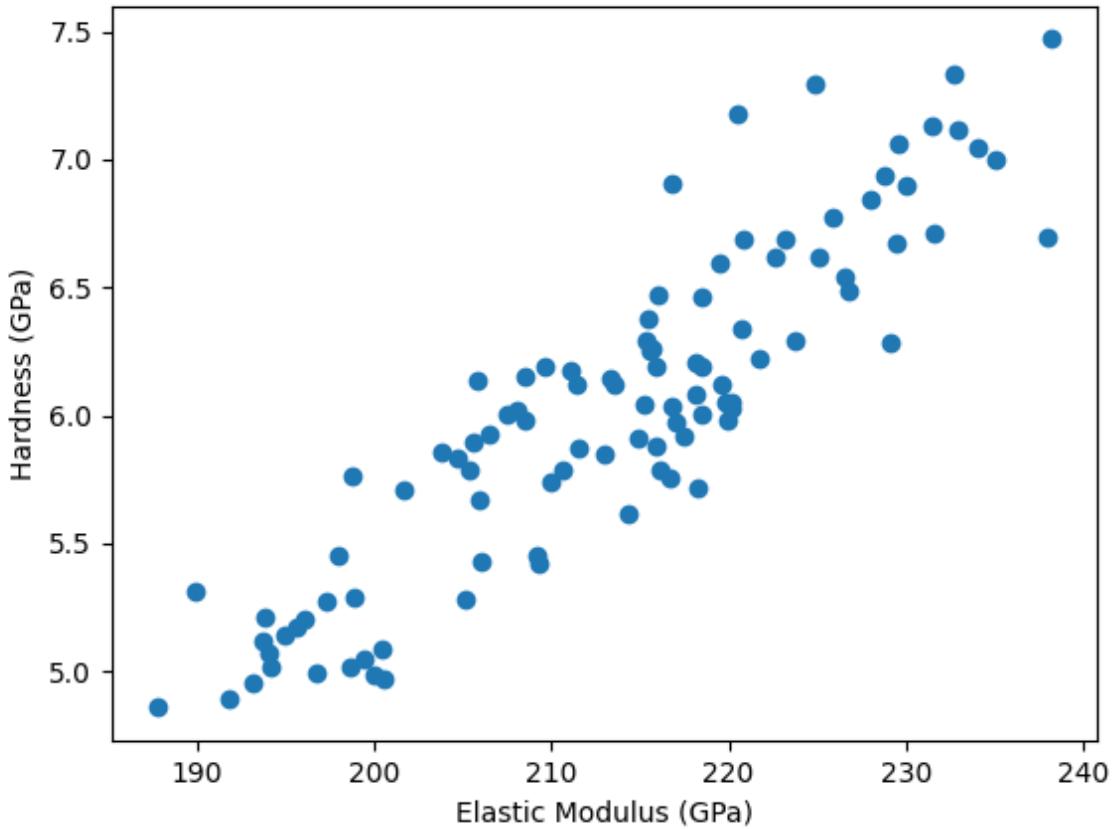
M P Gururajan, Hina Gokhale and N N Viswanathan

Department of Metallurgical Engineering & Materials Science
Indian Institute of Technology Bombay

September, 2024

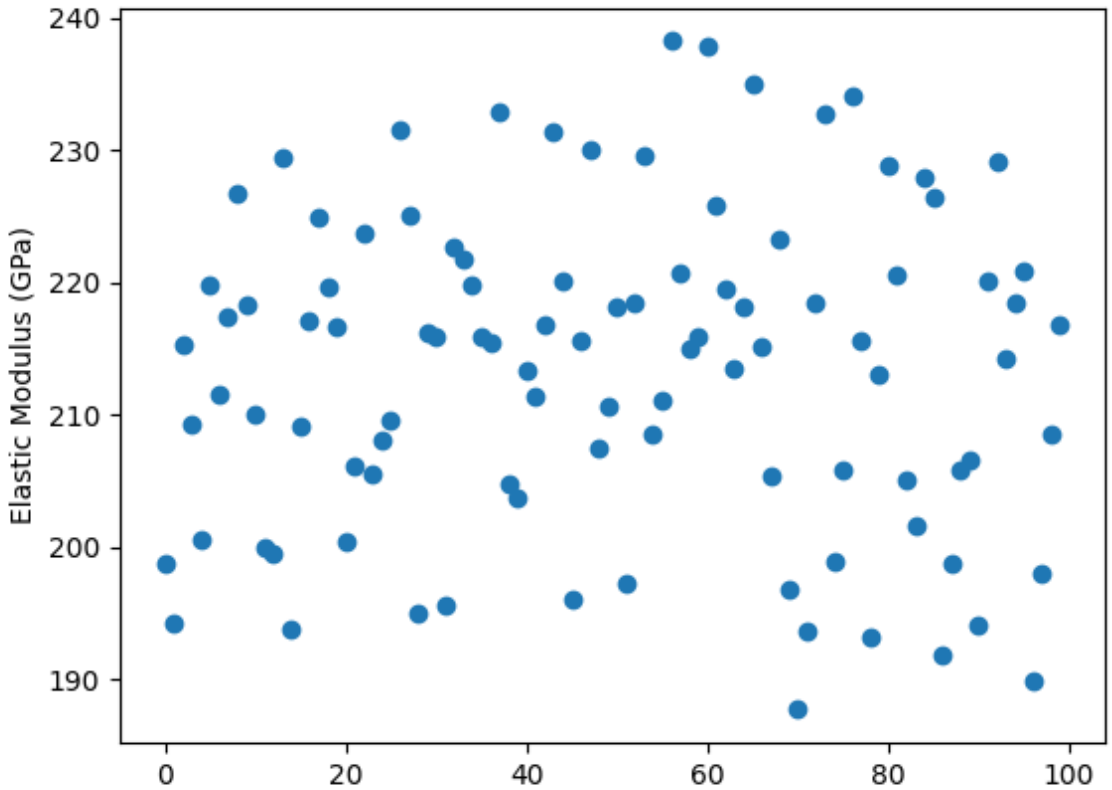
Outline

Nanoindentation: modulus - strength



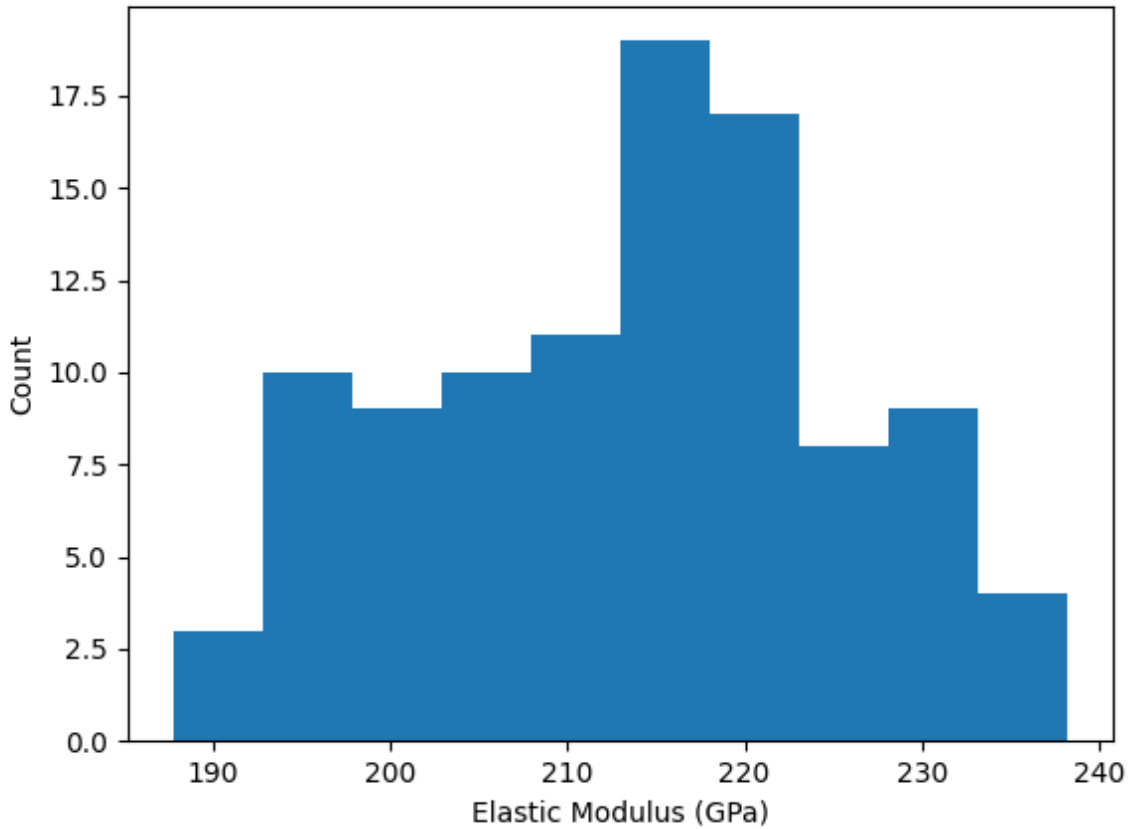
Data courtesy: Mr Subhas Bhunia

Modulus: data



Data courtesy: Mr Subhas Bhunia

Modulus: Histogram

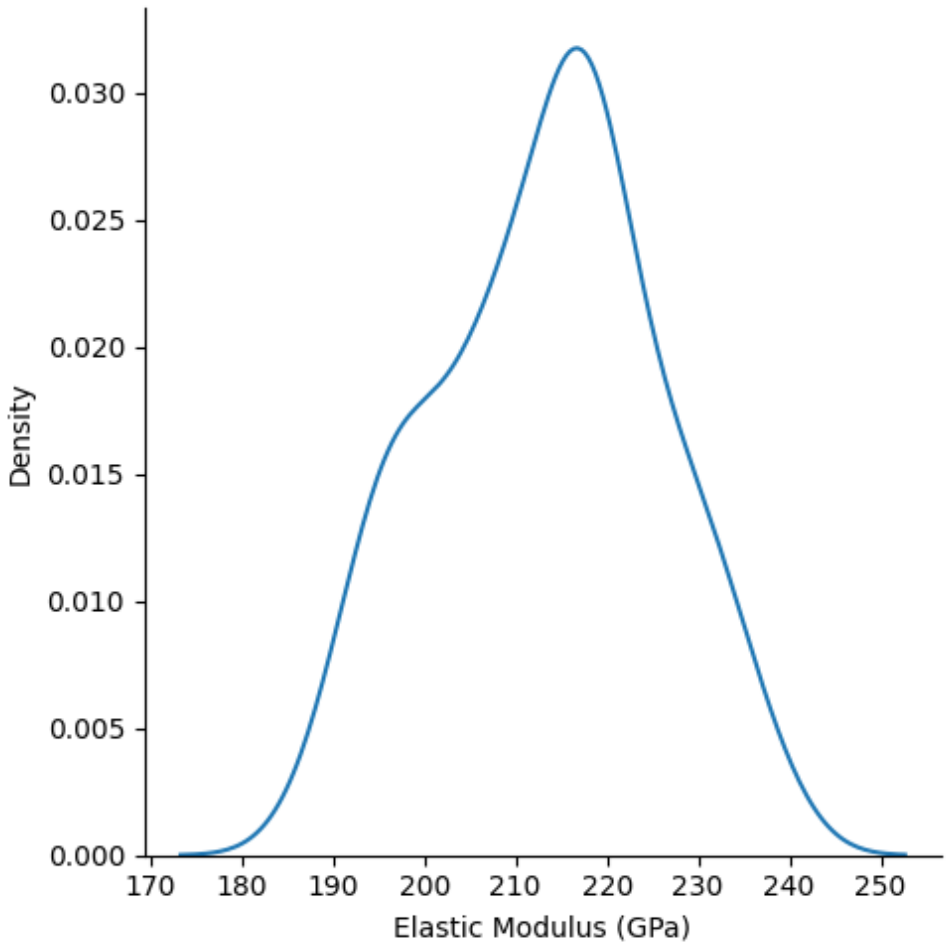


Data courtesy: Mr Subhas Bhunia

Probability density

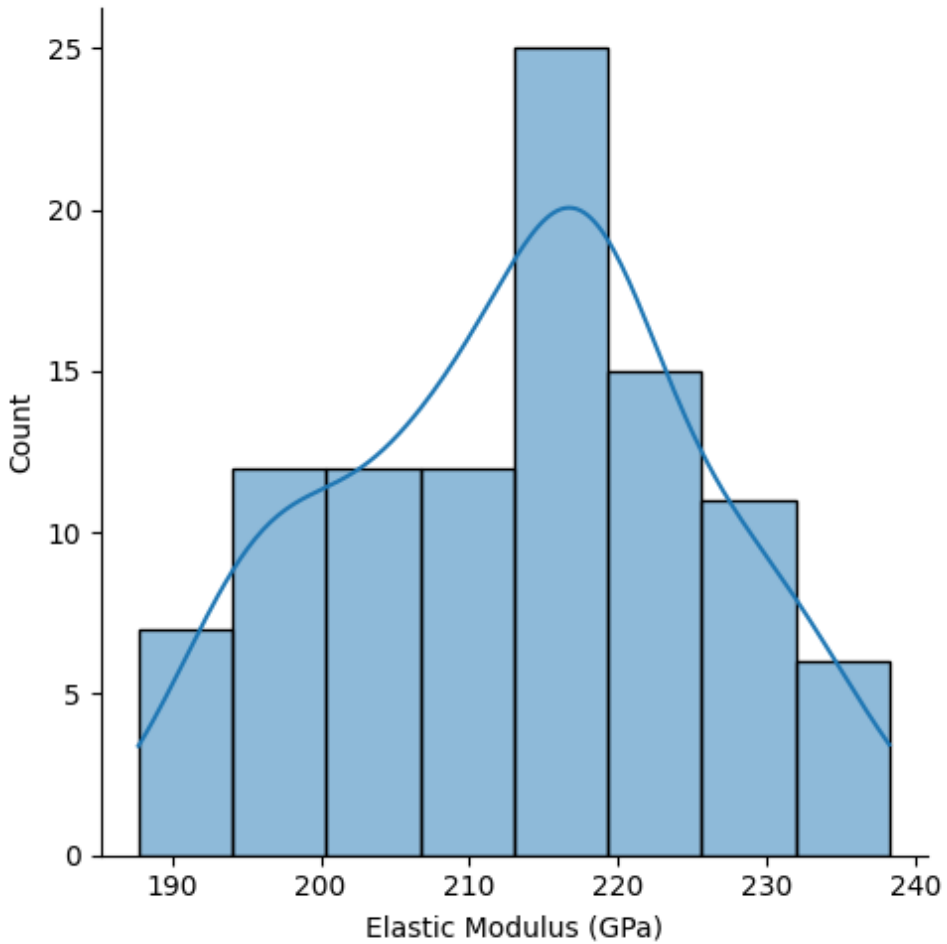
- Not normalised histogram
- Density: continuous variable!
- `seaborn` library to obtain density plots

Modulus: Density plot



Data courtesy: Mr Subhas Bhunia

Modulus: Density and histogram plots

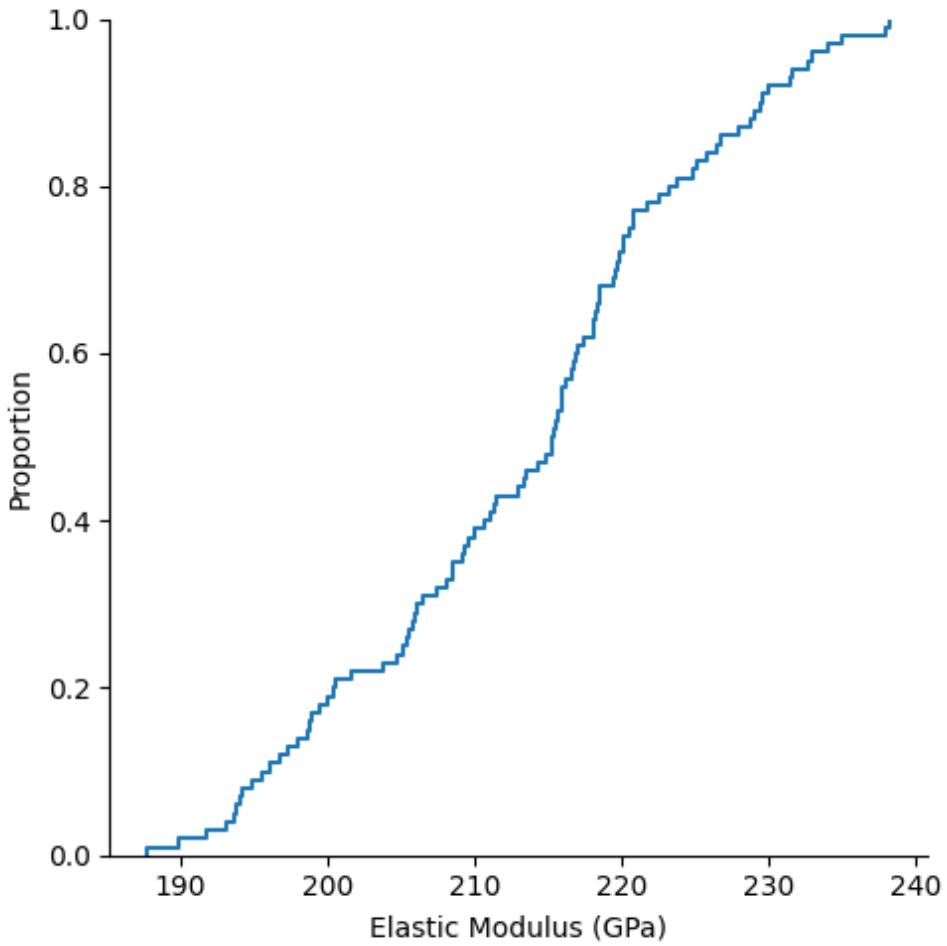


Data courtesy: Mr Subhas Bhunia

Empirical CDF

- CDF: Area under the density curve to a given point
- Remember: Particle size distribution in mineral processing (sieve analysis)

Modulus: Empirical CDF

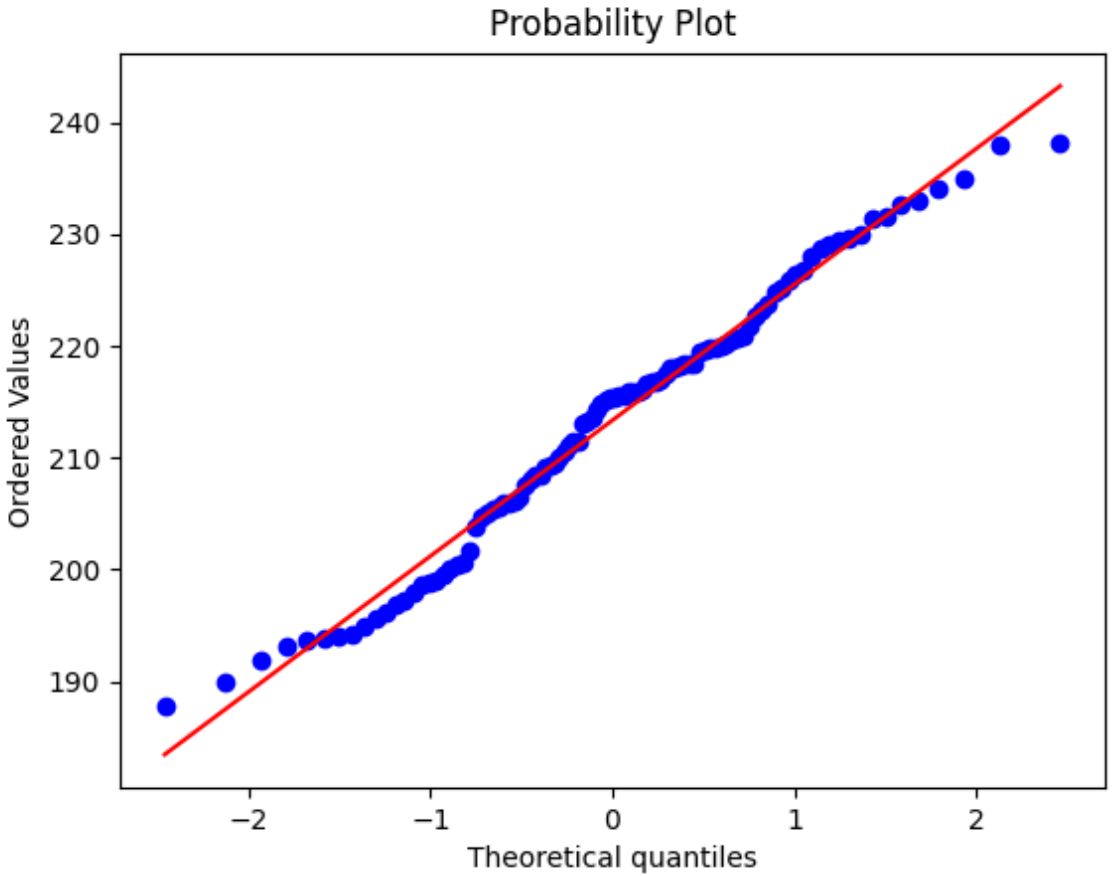


Data courtesy: Mr Subhas Bhunia

Q-Q plot

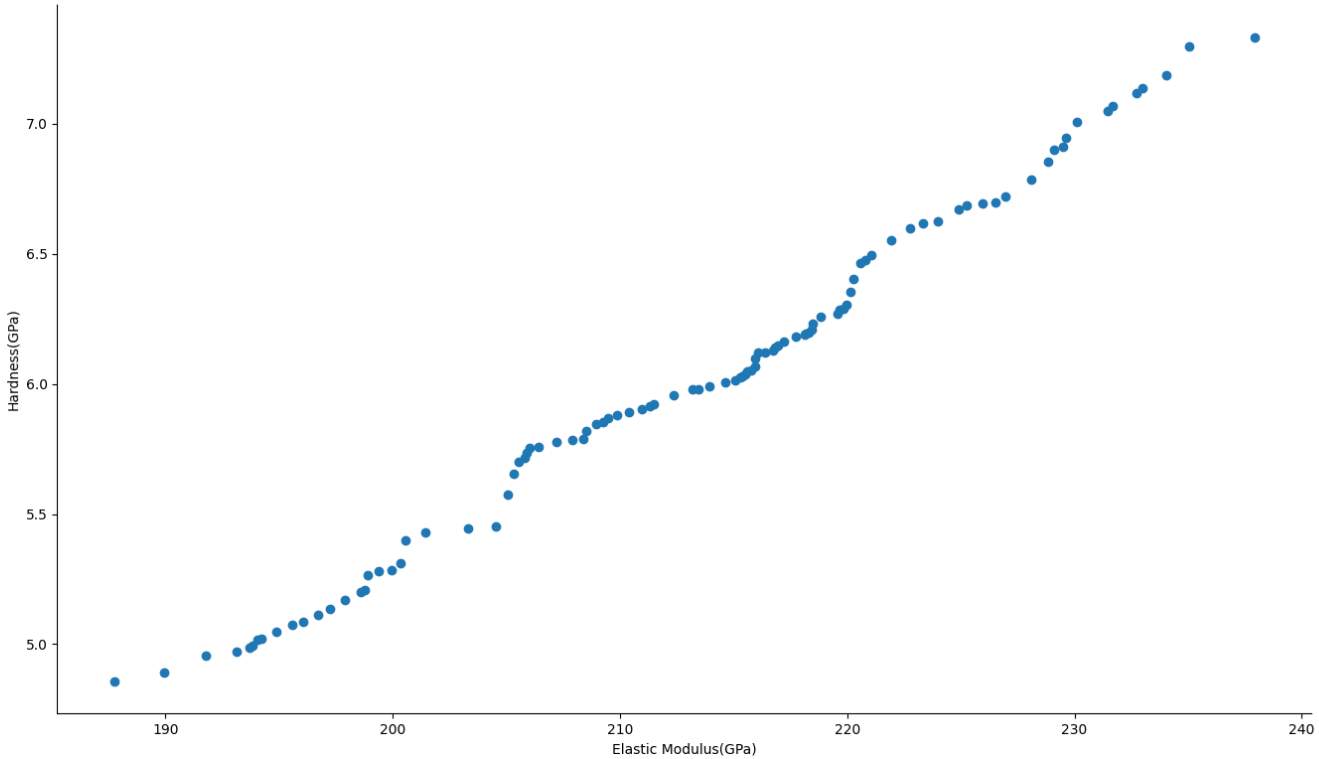
- Quantile: Range of probability distribution is divided into parts with each interval having the same probability
- Check the quantiles of the data with that of theoretical quantiles
- Deviations / agreement: tells whether the given data follows the distribution or not!
- Note: Q-Q plots can also be used to compare two data sets!

Modulus: Q-Q plot



Data courtesy: Mr Subhas Bhunia

Modulus and Hardness: Q-Q plot

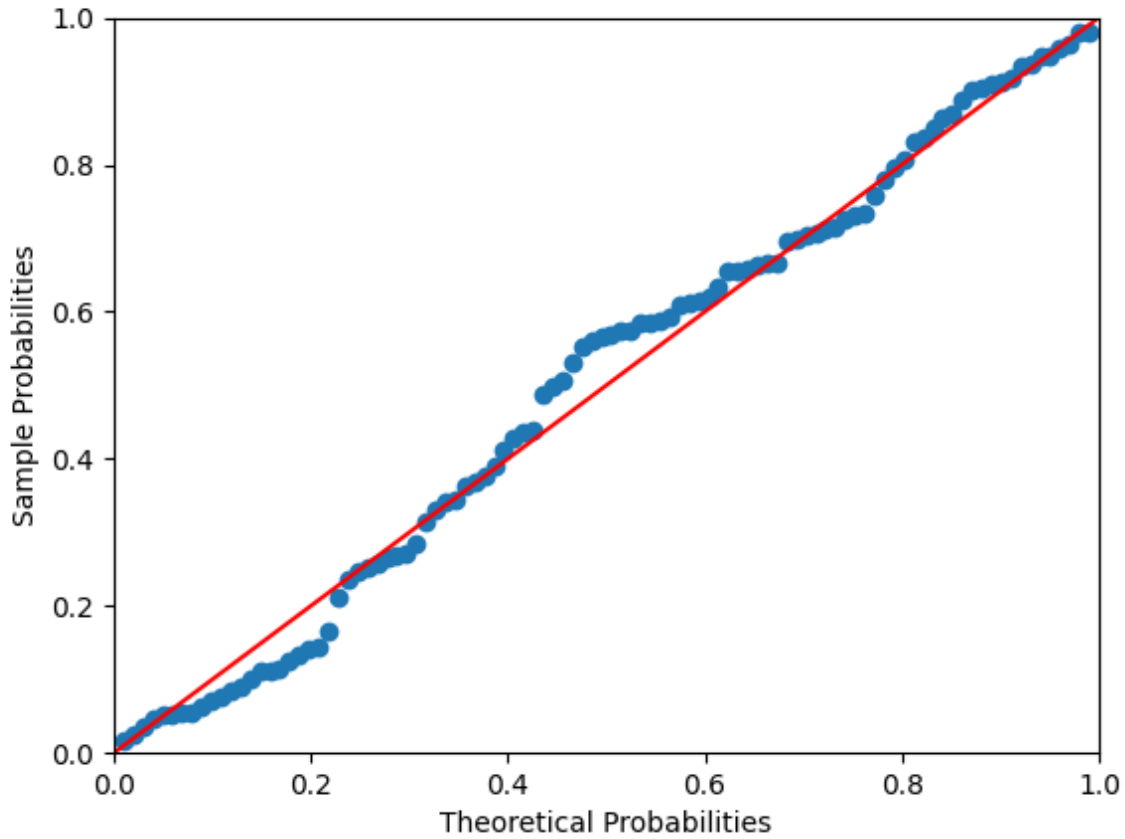


Data courtesy: Mr Subhas Bhunia

P-P plot

- Also to compare data with theoretical distribution
- Theoretical proportion versus actual proportion

Modulus: P-P plot



Data courtesy: Mr Subhas Bhunia

Reference

- Chapter 28: Smoothing of *Introduction to Data Science Data Analysis and Prediction Algorithms with R* by Rafael A. Irizarry
- Class notes used in the HarvardX Data Science Series.
- A free PDF (of the October 24, 2019 version of the book) available
- The R markdown code used to generate the book: available on GitHub
- Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International CC BY-NC-SA 4.0.
- For announcements related to the book on Twitter: follow @rafalab.

Data smoothing

- Data smoothing: important concept in ML
- Curve fitting, low pass filtering: other names
- Extract trend from noisy signal: smoothing
- Why is this useful?

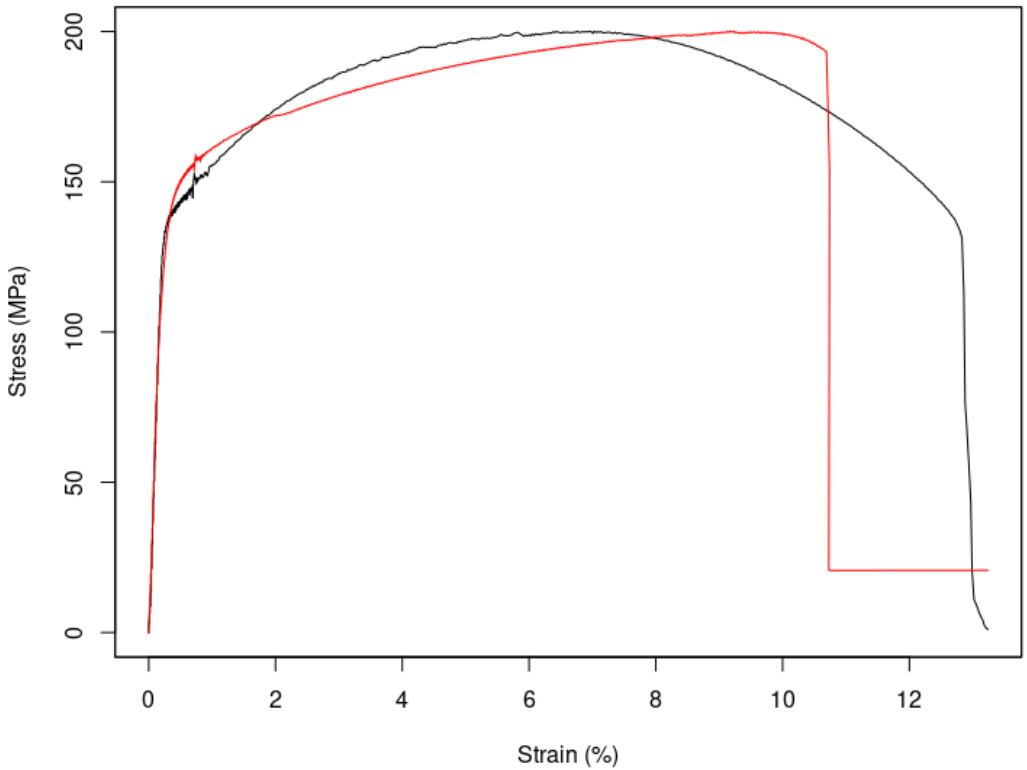
  ...the concepts behind smoothing techniques are extremely useful in machine learning because conditional expectations/probabilities can be thought of as trends of unknown shapes that we need to estimate in the presence of uncertainty.

–*Introduction to Data Science Data Analysis and Prediction Algorithms with R* by Rafael A. Irizarry

Noisy data

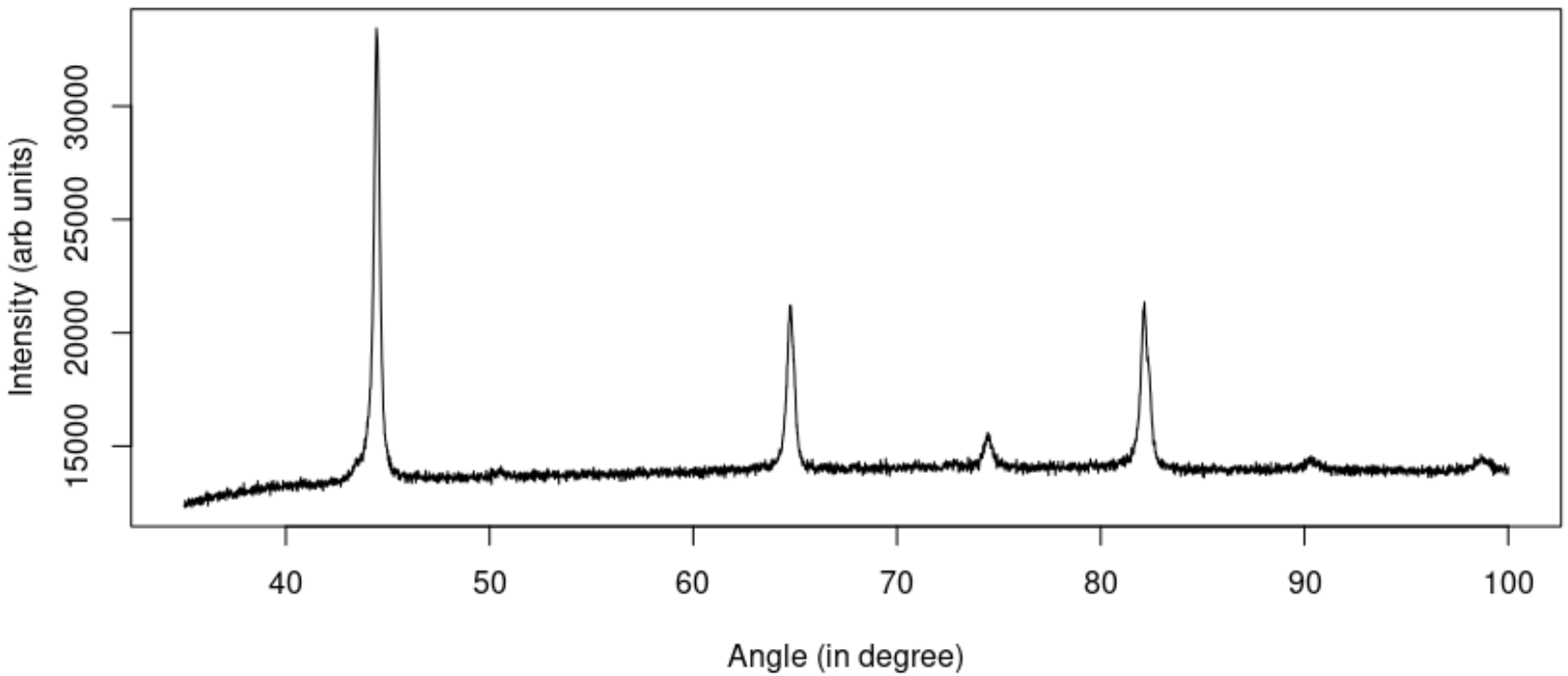
- Ubiquitous in metallurgy / materials science!
- Stress-strain data
- XRD pattern
- Simulation data: MD order parameter, for example

Stress-strain curves



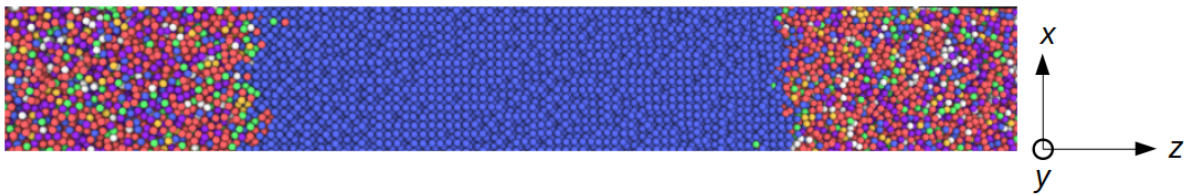
Data courtesy: Professor Prita Pant

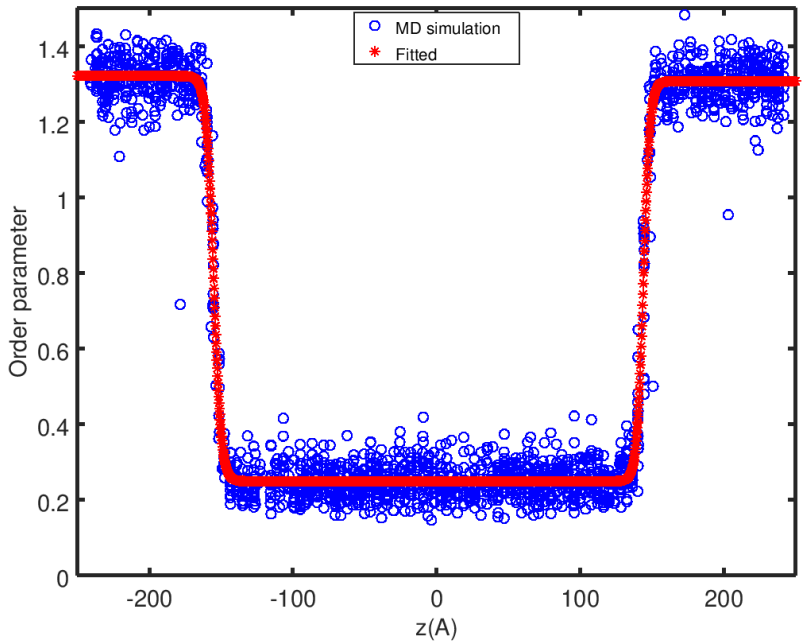
XRD pattern



Data courtesy: Mr Subhas Bhunia

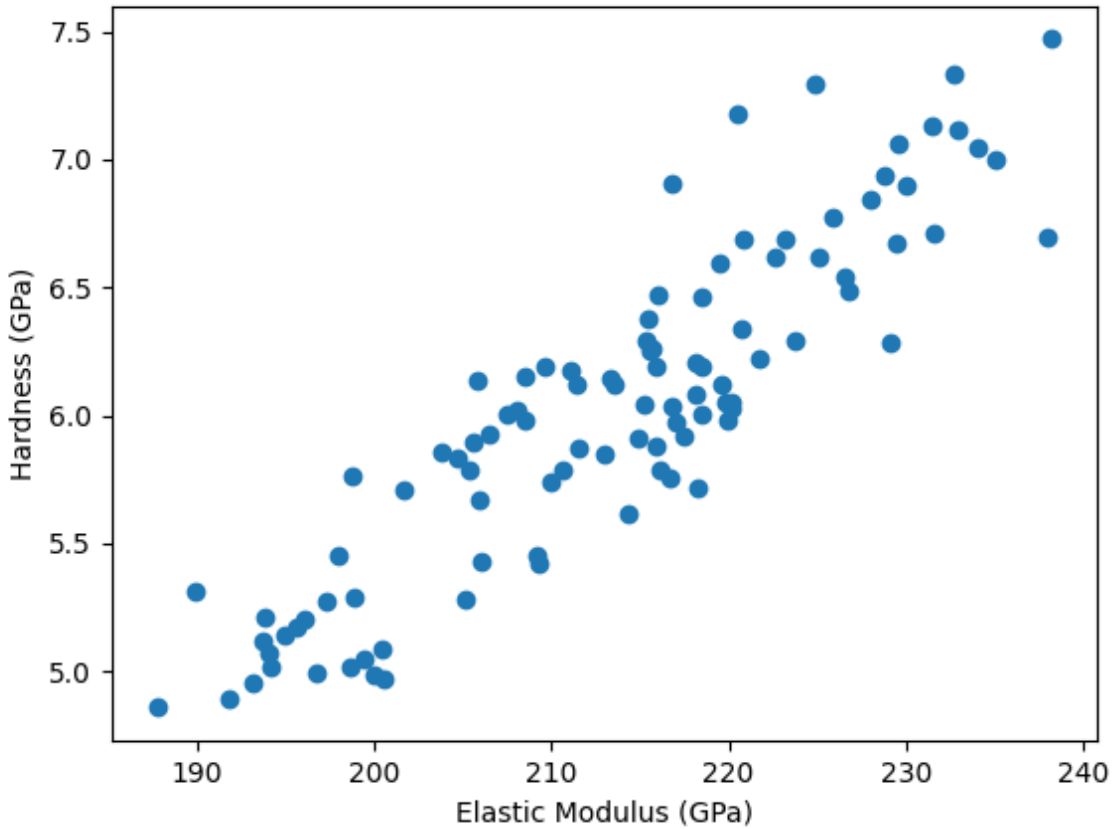
MD: order parameter





Data and image from Sushil Kumar et al, http://arxiv.org/abs/2105.14521

Nanoindentation: modulus - strength



Data courtesy: Mr Subhas Bhunia

Smoothing

- Take a window and average; move the window
- Moving average!
- Bin the data and take bin average! Mean bin method
- Fit a line; fit for entire data set or for local regions
- Can we use the same techniques to smooth XRD data?
- What happens to peak positions when we use moving average?
- What happens to smaller peaks when we use moving average?
- How to deal with such cases?

Data cleaning

- Smoothing: just one aspect
- Reset zeros: shifting
- Analyse leaving out certain data points – but without compromising data quality

Thank You!

Questions, clarifications, comments?