Data Analytics for Materials Science

A.D. (Tony) Rollett

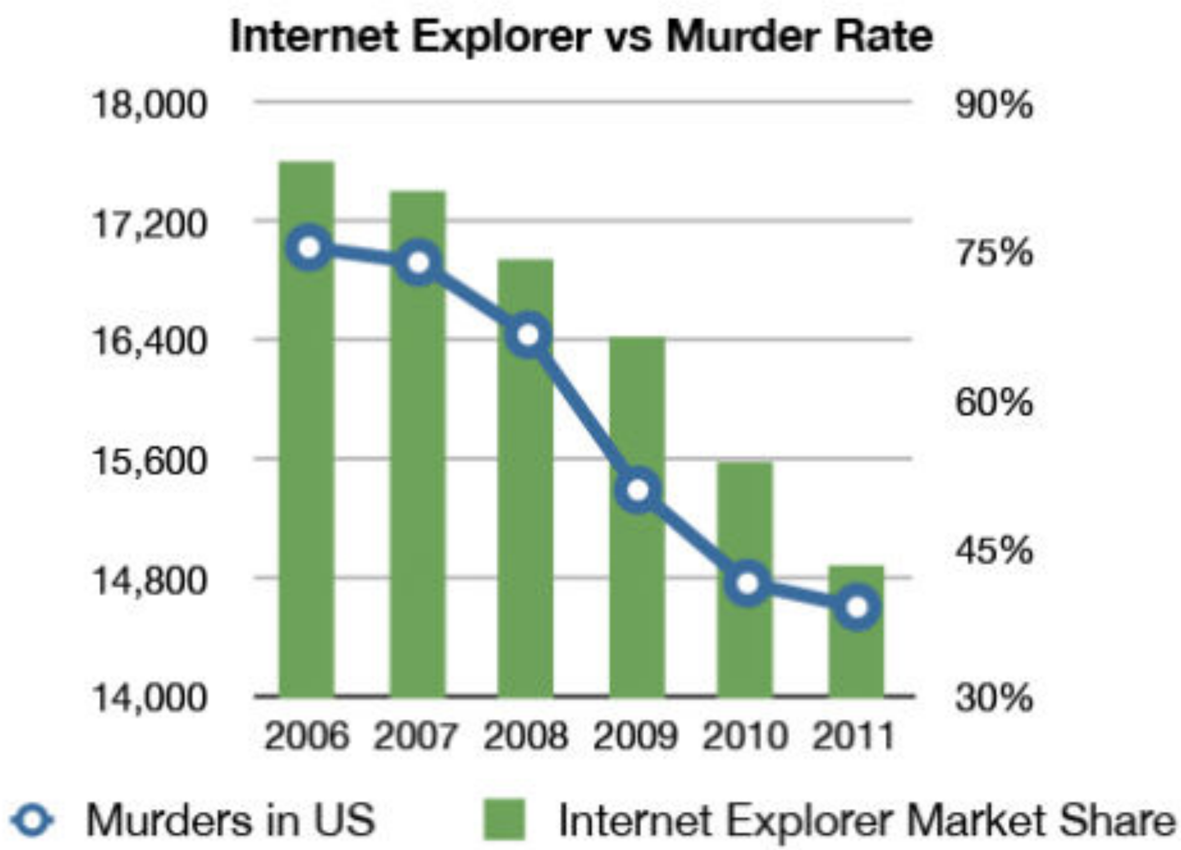
Dept. Materials Sci. Eng., Carnegie Mellon University

Introduction

Lecture 1

Revised: 3rd Feb. 2021

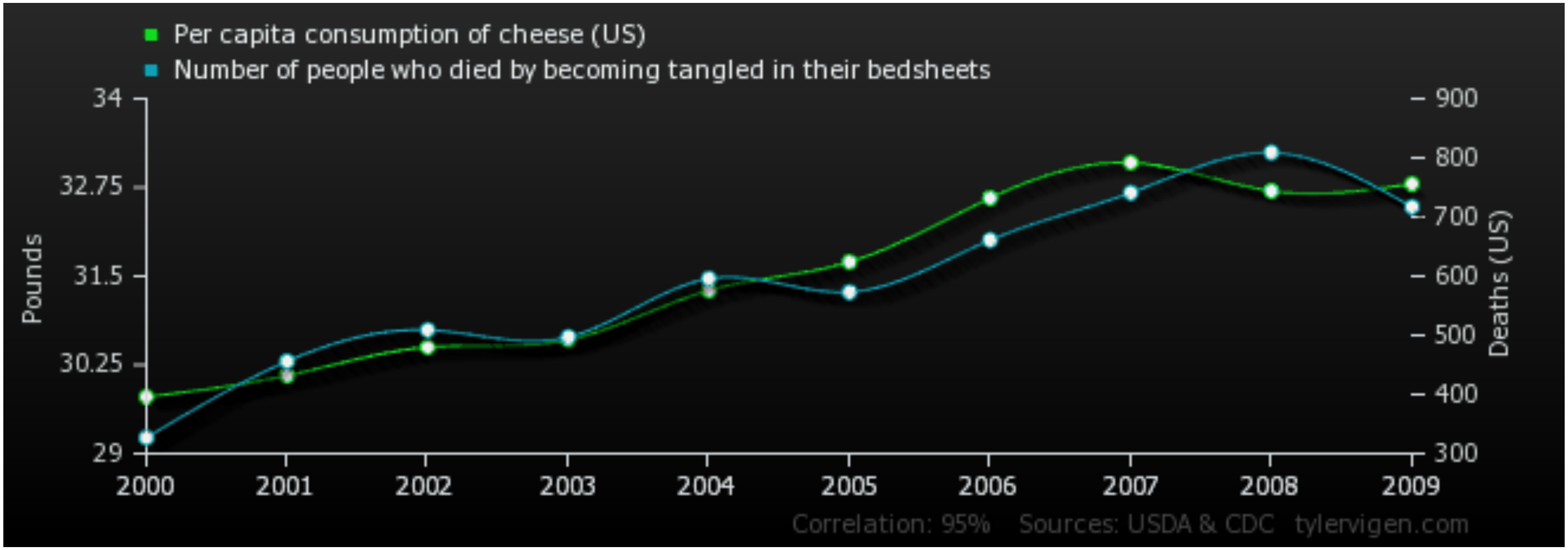
Statistical correlation and causation



As much as one may dislike Microsoft software, there is probably **not** a causal relationship.

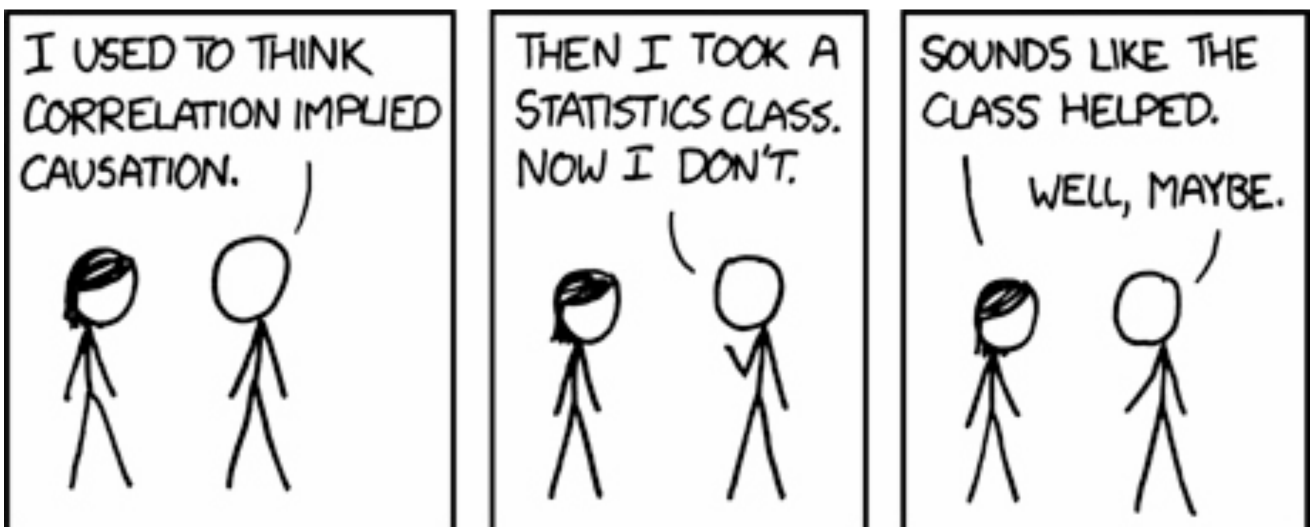
http://gizmodo.com/5977989/internet-explorer-vs-murder-rate-will-be-your-favorite-chart-today

Correlation ≠ Causation



Correlation doesn't imply causation!

Data points out the correlations.
But, physics should be used to decided causality!



Source: xkcd

Course Objectives

The main objective of this course is to introduce you to the use of *data analytics* in materials science.

Data Analytics is a very large subject. In principle, serious students can teach themselves from books, on-line materials, on-line videos etc. However, most of us find it easier to do this with examples drawn from current practice in our own field of study, i.e., materials science & engineering. Furthermore, the choice of topics is so large that it is helpful to see which ones have proven value in terms of recent research.

Topics

The topics will include:

- Examples of distributions found in materials science, e.g., log-normal, similarity of distributions, and "p values"
- Linear Regression, Correlation
- ANOVA
- Multiple Linear Regression with feature selection
- Principal Component Analysis (PCA)
- Canonical Correlation Analysis (CCA)
- Decision Trees: Regression vs Classification
- Random Forest: Ensemble Methods
- Neural Network
- Computer Vision
- Databases for materials data
- Symbolic Regression with Bingo

References

- Applied Multivariate Data Analysis: Regression and Experimental Design Vol. 1, J.D. Jobson
- Applied Multivariate Data Analysis: Vol. 2: Categorical and Multivariate Methods (Springer Texts in Statistics), J.D. Jobson

– the two Jobson volumes are available electronically from the CMU Library

Interesting recent article:

https://towardsdatascience.com/getting-started-in-materials-informatics-41ee34d5ccfe

Test, Exams, Grading Policy

This is also provided in the Syllabus; check for inconsistencies!

Homeworks: max. of 1 per week, variable points

Examinations: two in-class tests, see weighting below; also a Final

Grading Policy	A > 90 %
	B > 80 %
	C > 70 %
	D > 55 %

The instructor may give an Oral exam in borderline cases.

Weighting:	Homeworks	30
	Tests (2 take-homes)	20
	Final	25
	Independent Study	20
	Participation	5

The two in-class tests are intended to provide a check on how well you are learning the material. The *independent study* will most likely involve finding a dataset by yourself and making your own decisions (to be documented and justified) about how to apply data analytics. Last year (2020), we provided a dataset about to those who wanted one.

General Policies

Open door, office hours: if I am in my office, I am available. I find it difficult to offer fixed office hours, so it is best to email questions to me and request a meeting if my response is insufficient.

Accommodations for Students with Disabilities: if you have a disability discuss it with me as soon as possible.

Your State of Mind: take care of yourself! Call CaPS: 412-268-2922 or (if life-threatening) Campus Police: 412-268-2323.

Your One and Only Warning: Zero Tolerance of Cheating & Plagiarism

Class Presence and Participation: class presence and participation points are given to encourage your active class participation and discussion, especially for graduate students in the journal club session at the end of each class. You will be rewarded with a perfect score as long as you frequently come to class and actively contribute to the class discussion during recitations and lectures. It is polite to inform the instructor if you know that you will not be able to attend class.

Informatics

A field of research in which

- information science, processing, and systems combine to examine the structure and behavior of information (a Wikipedia definition)

New ways to extract information from data:

- data can be of mixed type and mixed quality
- enables integrating modeling "data" with experiment
- may play critical role in "linking length scales"

Caveats:

- data must be organized and represented properly
- quality of information depends on data

Informatics does not answer "Why?"

- provides places to look for "why"

The Promise of Materials Informatics

“All models are wrong, but some are useful”
- George E.P. Box, 1976.

Materials informatics facilitates “accelerated insertion of materials into engineering systems and rapid multiscale design and optimization of materials properties.”
- Krishna Rajan (University of Buffalo), 2006.

“AI is a basic technology that can accelerate the pace of materials discovery and clean energy”
- Eric Krotkof (COO, Toyota Research Institute) , 2017.

Microstructure Informatics

“Microstructure Informatics addresses the challenge of linking material’s internal structure (i.e. microstructure), its evolution through various manufacturing processes, and its macroscale properties (or performance characteristics) through the development of data science algorithms and computationally efficient protocols."

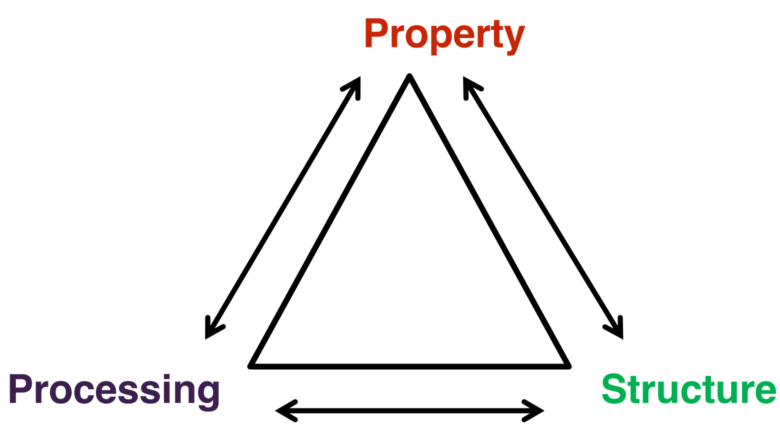
Microstructure Informatics Using Higher-Order Statistics and Efficient Data-Mining Protocols, Kalidindi et al, JOM, April 2011

Less formally stated: it is a common experience in our subject that we acquire multiple types of quantitative data. At the highest level, we seek explanations and quantitative models for the causation of phenomena (science) with the aim of controlling and exploiting those phenomena (engineering). We use hypothesis testing to identify physical mechanisms and we often plot one quantity versus another to explore relationships. However, it is increasingly common to have multiple types of data from a single source. For example, microstructures often yield grain size, grain shape and orientation (3 numbers), which yields a spreadsheet with 5 columns and as many rows as there are grains.

Either explicitly or implicitly, we often have inputs and outcomes, or, independent variables and dependent variables. These sets of variables may be linked, i.e., correlated, which suggests that it is reasonable to play with different combinations of the measured quantities in order look for connections. This is precisely what correlation analysis can assist with. There are many other analytical techniques that may also be useful.

Materials Informatics

- Incorporation of data mining, machine learning (ML), computer vision (CV) and artificial intelligence (AI) into materials science and engineering
- Inspired by Cheminformatics and Bioinformatics (however, not as common)



- **Microstructure informatics:** Identify trends and correlations in parameters like grain size, shape and crystallographic texture etc.
- **Process modeling:** Goals are descriptive, prescriptive & explanatory.
- **Property databases:** Quantifications of property responses & their modeling

Applications of materials informatics:

- Descriptive modeling: probability distributions of structure parameters, tails of stress-strain distributions, variants clustering, causes of abnormal grain growth
- Predictive modeling: constitutive models for extrapolation, predicting stress hotspots, powder bed additive process monitoring, hot spots from rough surfaces
- Discovering patterns and rules: correlation analysis
- Classification: predicting microstructure image classes, classifying powders, classifying fracture surfaces

Discovering patterns and rules: recursive partitioning

Objective:

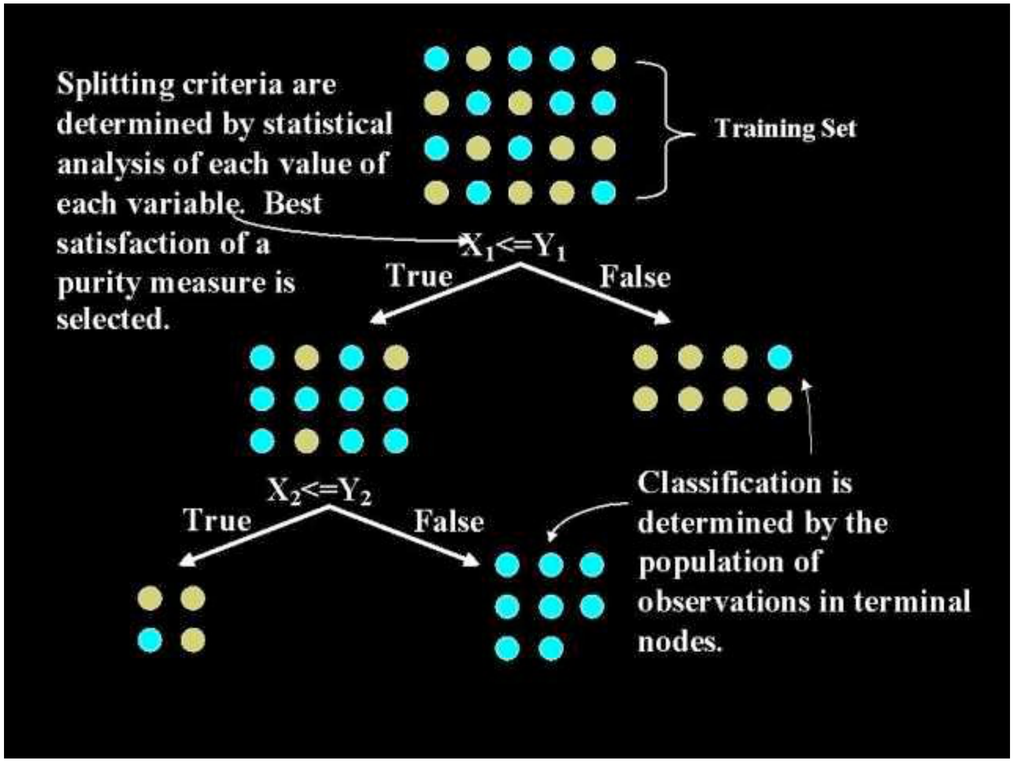
to develop prediction and classification rules
⇒ a decision tree

Method:

(1) split the data set by defining criteria

(2) check on "purity" of split

(3) define stopping criteria



http://ai-depot.com/Tutorial/DecisionTrees-Partitioning.html

Data mining: the tools of informatics

Tasks:

- *exploratory data analysis*
  - visualize what is in the data
- *descriptive modeling*
  - probability distributions
  - cluster analysis
- *predictive modeling*
  - classification and regression
- *discovering patterns and rules*
  - spotting behavior
  - outlier identification
- *retrieval by content*
  - match desired patterns

All these tasks employ a suite of methods

Deciding what method and how it can be used is the "art" in informatics

'Principles of Data Mining', Hand *et al.* (2001)

Exploring data: multidimensional data

Ashby maps are two-dimensional

- need many of them to describe data sets with many variables
- may need to determine multiple correlations (not visual)
- if p is the number of variables, have $p(p-1)/2$ pairs

We cannot visualize multidimensional maps for $p>3$.

- we can try to *reduce the dimensionality* of the data to make it more accessible for analysis
- numerous methods available for dimension reduction, including factor analysis, linear discriminant analysis, and *principal component analysis* (PCA) or *canonical correlation analysis* (CCA)

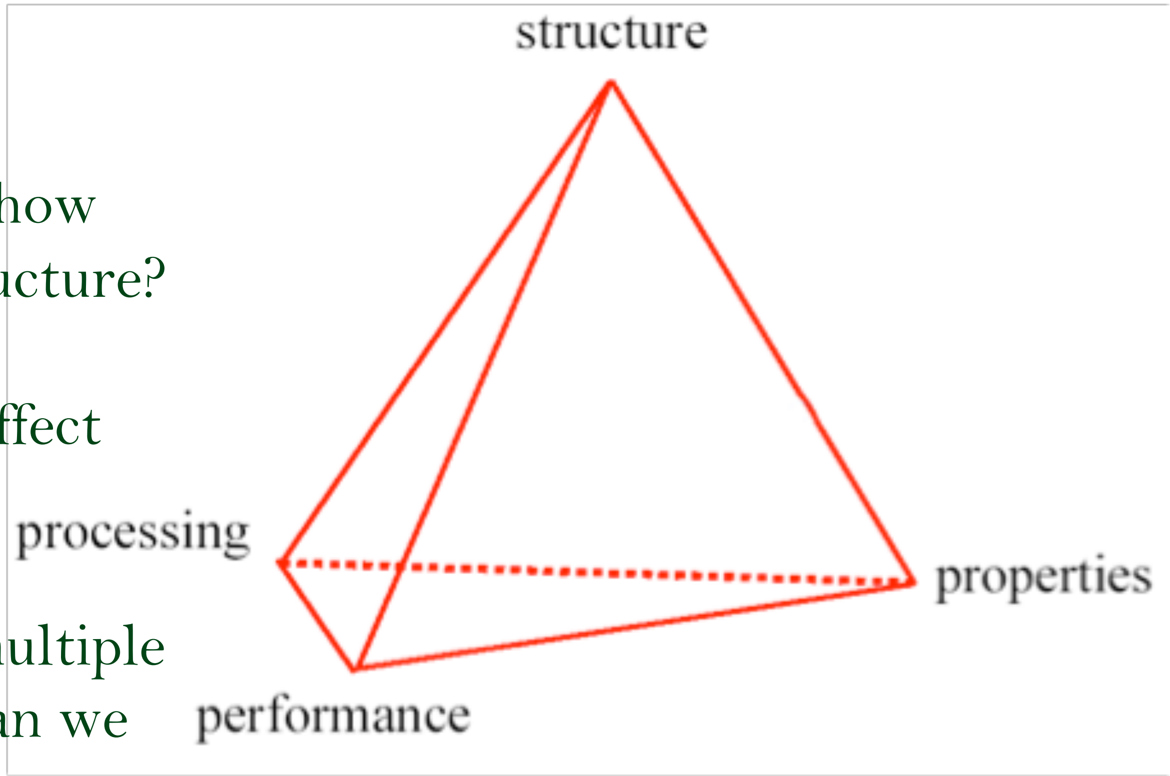
The Materials Science and Engineering Paradigm

The key questions:

How do we best understand how processing affects (micro)structure?

How does (micro)structure affect properties?

All four categories contain multiple variables/parameters: how can we analyze them all at once?!



‡Materials Science and Engineering for the 1990s – Report of the Committee on Materials Science and Engineering (Washington DC, National Academy Press, 1989).

Microstructural parameters

crystal structure
grain size
grain shape
grain orientations (texture)
phase structure
composite structure
porosity
dislocation substructures

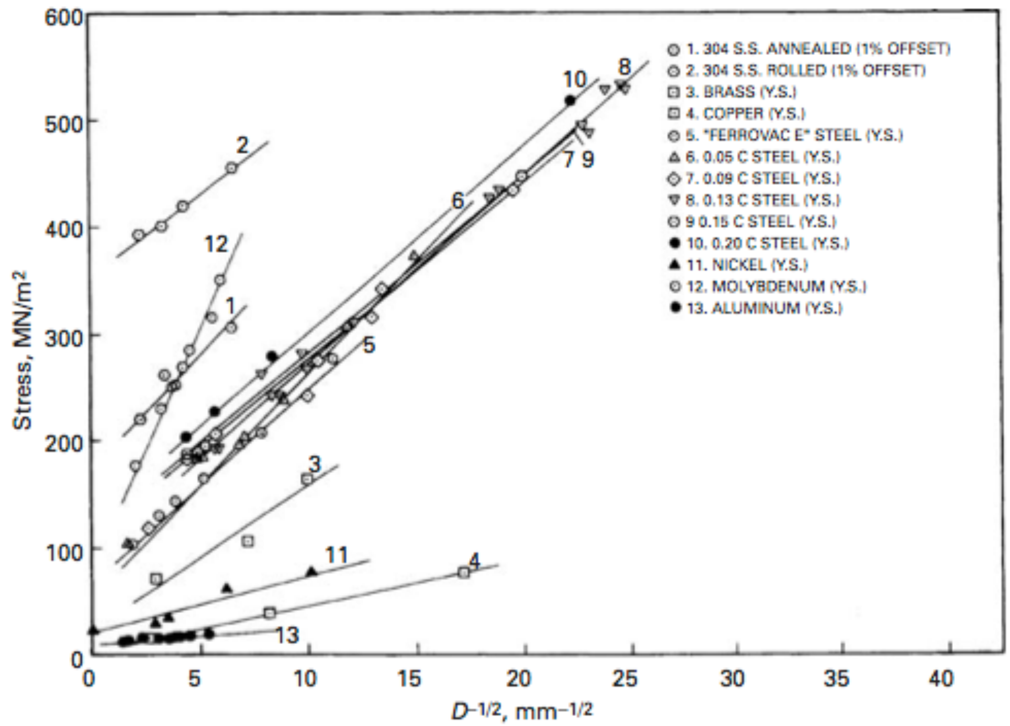
Material properties

stiffness
strength
toughness
formability
conductivity
corrosion resistance
magnetic susceptibility
dielectric constant

What aspects of microstructure affect properties?

Simple example: Hall-Petch relation

Experimentally-based relation between strength and mean grain size:



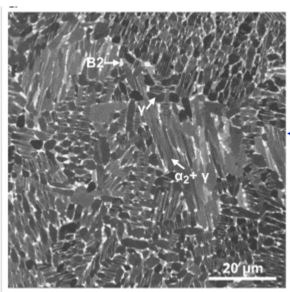
$$\sigma_y = \sigma_o + kD^{-1/2}$$

H-P relation ssing to obtain more desired properties *in spite of not having a totally accepted theory/model for why it seems to hold*

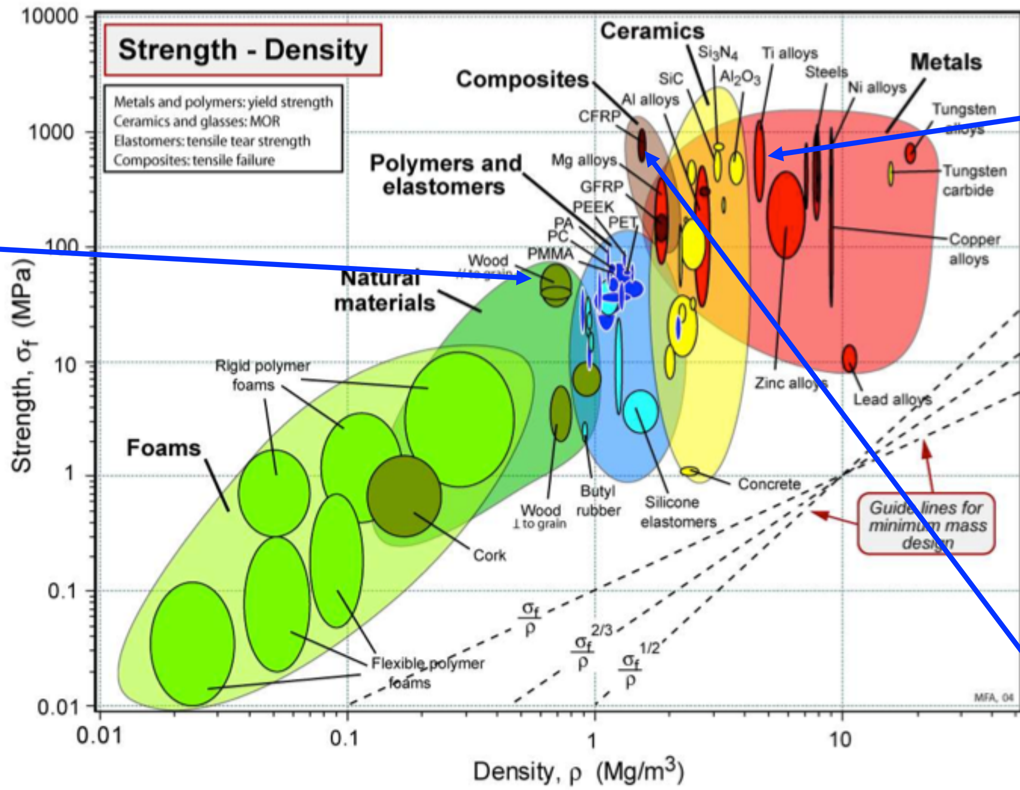
figure from Meyers and Chawla textbook

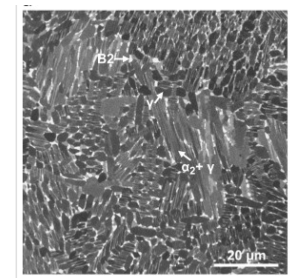
Two-variable property (Ashby bubble) maps

- pairwise correlation of properties - a simple version of "informatics"

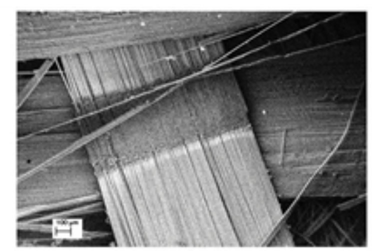


Ashby, Cambridge





Fraser, OSU



Goren, Zurich

Modern characterization tools, combined with advances in modeling, are leading to a deluge of data.

Tony Fast, Ga. Tech

Challenges in extracting information from materials data

Emerging ability to characterize 3D structures:

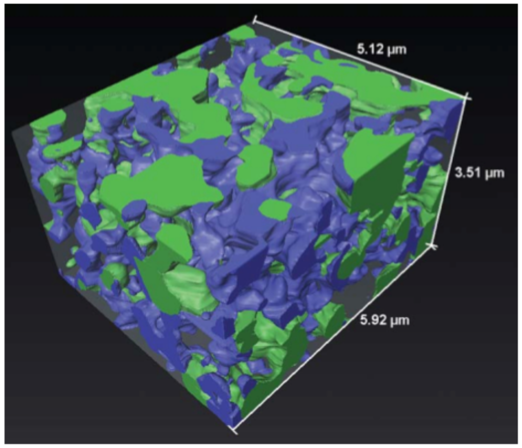
- representation of structures
- microstructure/property relations

Integration of data types:

- data with differing degrees of accuracy and precision
- modeling “data”
- new tools (atom probe, scattering, ...)

Images are often hard to interpret

We need new statistical tools to extract more complex interrelationships from data, e.g., computer vision methods, CNNs



Solid Oxide electrode: Ni (green), YSZ (gray), pore (blue). Kammer and Voorhees, MRS Bulletin 33, 603 (2008)

Acknowledgements

- **Boeing:** Nathan Ashmore, James Cotton, Karen Thacker and others
- **Air Force Research Laboratory (AFRL), Dayton OH:** Dr. Brian Gockel, Dr. Sean Donegan, Dr. Mike Groeber
- **BlueQuartz Software:** Mike Jackson
- **Indian Institute of Science Bangalore:** Prof. Dipankar Banerjee
- **Georgia Tech:** Prof. Hamid Garmestani, Prof. Stephen Liang, Jason Allen, Pan Zhipeng
- Michigan State University: Dr. Shanoob Balachandran
- Kumamoto University: Prof. Donald Shih
- **Carnegie Mellon University:** Prof. Liz Holm, Prof. Warren Garrison, Yu Feng, Vahid Tari, Shivram Kashyap, Dr. Samikshya Subedi, Dr. Tugce Ozturk, Ankita Mangal, Ross Cunningham, Saransh