27-737 Data Analytics for MSE: How to Choose an Approach

Anthony Rollett & Amit Verma

Materials Sci. & Eng., Carnegie Mellon University, Pittsburgh, PA.

Revised:
Apr. 21st, 2021

What is the End Goal?

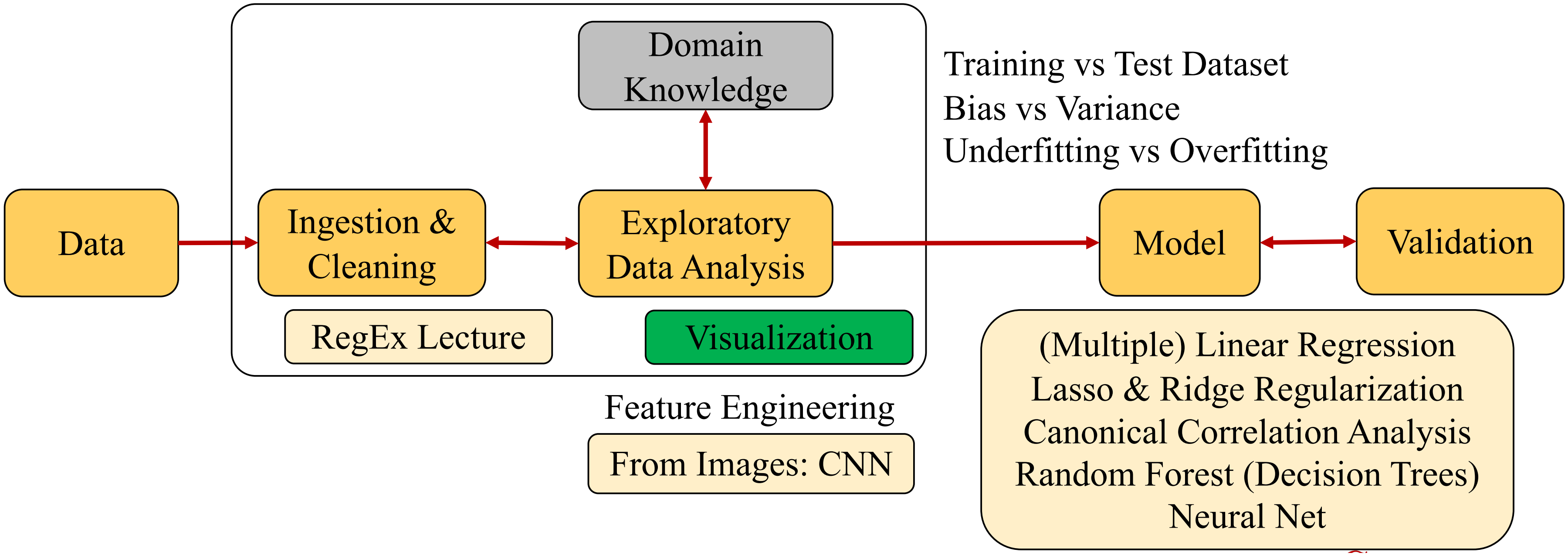
- Ask yourself what you want to get out of the analysis. Do you have a well-formed hypothesis, or are you exploring a technique and simply accumulating results?
- For example, if you are investigating the Hall-Petch effect in metals, you probably already have a precise idea of what the expected result is because of the large body or prior work. In this circumstance, it is unlikely that your data will lead to a new idea and your task is to apply data analytics to the particular material or testing method at hand.
- Since (yield) strength is the main outcome of varying grain size (or layer thickness), one would most likely apply multiple linear regression with tests of statistical significance for the influence of other variables to see if they contribute to the variability (in strength).
- Traditionally, we would assume the √d dependence, draw lines through the data and evaluate by eye. E.g., “Strength of Nanoscale Metallic Multilayers”, S. Subedi, I.J. Beyerlein, R.A. LeSar, and A.D. Rollett, *Scripta mater*. (2017).
- Note that assuming a physical model precludes the application of statistical tests for validity. Of course, one can and should evaluate standard deviation, confidence bounds etc.

What is the End Goal? 2

- At the other end of the spectrum, consider the analysis of micrographs (of cross-sections of material). This started with the question of whether one can train a computer to recognize different kinds of materials. This boils down to a classification exercise and is therefore very similar to classifying images into dogs, cats, buses, cars, trucks, trains etc.
- The challenge is to find methods for extracting information from each image that leads to a statistically useful descriptor of each image, which one can call "*feature extraction*". More importantly, the feature vector, however obtained, must allow one image type to be distinguished from another.
- Until a few years ago, feature extraction was based on explicit identification of specific features such as spots and lines. This can be found, e.g., in the earlier work at CMU on powder classification: “Computer Vision and Machine Learning for Autonomous Characterization of AM Powder Feedstocks”, DeCost *et al*., *JOM*, **69** 456–465 (2017).
- More recently, convolutional neural nets have proven to be very powerful, and certainly more effective than the older techniques. A good example is the classification of defects provided in the Northeastern University database.

Flow Chart

- Ultimately, we aim to develop a decision tree about how to analyze data. The example provided earlier shows the beginnings of how to go about this. The early stages should focus on exploring and visualizing the data. Once you make some decisions about potential models, then try to obtain quantitative results.


Domain Knowledge
Training vs Test Dataset
Bias vs Variance
Underfitting vs Overfitting
Data
Ingestion & Cleaning
Exploratory Data Analysis
Model
Validation
RegEx Lecture
Visualization
Feature Engineering
From Images: CNN
(Multiple) Linear Regression
Lasso & Ridge Regularization
Canonical Correlation Analysis
Random Forest (Decision Trees)
Neural Net

What is the End Goal? 3

- A different circumstance is provided by data stemming from experiments in which a researcher seeks to identify the causes of variations in a certain signal.
- For example, in the application of CCA to the measurement of cathodoluminescence in CdInTe, along with EBDIC as a function of grain boundary type, the reason to adopt this approach was that a) direct testing of the hypothesis that these properties were related to GB character failed and b) no physics-based models were available. Therefore, it seemed worth applying CCA to find out if we had missed anything. This can be construed as an open search for a model, albeit confined to linear relationships. See: "Grain-boundary character distribution and correlations with electrical and optoelectronic properties of CuInSe2 thin films", Abou-Ras et al., *Acta Materialia* **118** 244–252 (2016).
- CCA was useful because there were multiple input variables, compounded by the various ways in which grain boundaries can be represented.
- The results showed that there is indeed a relationship between CL and GB character albeit of a highly non-linear nature. They also showed that the CL and EBIC signals were not strongly correlated. Later on, a refinement of the analysis in "Data Analytics using Canonical Correlation Analysis and Monte Carlo Simulation" Rickman et al., *Computational Materials* **3** 26 (2017), showed that a stronger correlation could be obtained by allowing for non-linear (polynomial) relationships to be used.

Linear Regression

- The easiest model to develop and easiest to understand
- The classic example is the Hall-Petch relationship.
- Wait a minute: was this *really* linear regression?
  Discuss …
- Even linear regression has some hidden traps. E.g.,
  When we calculate the error as,
  MSE = $\sqrt{\{1/N \text{ sum}_i (y_i\text{-yhat}_i)^2\}}$,
  what are we assuming about the source of the variance?
  Discuss with sketches …
- All that said, when developing a model to explain data, the yhat versus y(measured) should be a straight line with slope=1 and passing through the origin, regardless of the linearity of y(x).
- You should *always* try linear regression first.

Multiple Linear Regression

- More common in real life is the situation where the outcome of interest depends on multiple factors.
- Here, we have to allow the model to include all possible inputs.
- Why can't we plot y(x) versus x? To state the obvious, the reason is that there are multiple "x"s. The practical limit is plotting a surface of z (dependent) versus (x,y). Or, if armed with 3D visualization (e.g., paraview), contour surfaces (of the response value) in 3D.
- We can, however, plot partial dependence, which we looked at briefly, which is equivalent to the partial derivative of yhat w.r.t. a single variable.
- For this situation it is always a good idea to try MLR before anything else, even if you suspect that a non-linear relationship will result in a model with lower loss.
- A tableau of correlation plots is always advised, both to visualize which inputs are most closely related to the output of interest, and to check for correlations among the inputs.
- Checkpoint: why might MAE be less sensitive to outliers than MSE? Discuss …

Advanced MLR

- What do we do when the number of inputs (columns) is (far) larger than the number of datapoints (rows)?
- The answer is that we have to determine which subset of the features/columns gives a model that can best explain the outcome of interest, i.e., minimizes the loss.
- Although this is always important, this is where cross-validation becomes useful so as to avoid over-fitting.
- The other useful concept here is that of *model complexity*.
  Discuss …
- Remember that using a model with more coefficients than datapoints is an under-determined model.
- There is (almost) an infinitude of methods. Let's focus on three methods: best subset, ridge regression and lasso.
  Discuss …

- *Best subset* has the attraction of being easy to understand and a plot of loss versus the feature count used serves to visualize what choice one is making.
- *Ridge regression* is the first stage in varying the weight (importance) attached to each input but cannot help where you have too many inputs.
- *Lasso* has the attractive feature that less important features are completely eliminated (set to zero) so that one can visualize the ranking.

Principal Component Analysis (PCA)

- Another classic linear method is PCA.
- It is generally presented as a method for dimensionality reduction. One rotates the data by performing an eigenanalysis and then inspects to see how few components (i.e., eigenvectors) are needed to explain, say, 90 % of the variance. Ideally, only 2 or 3.
- As we have started to see, however, sometimes one does not get a nice straight line in whatever PC space one uses. How then does PCA help us with fitting versus clustering?
  Discuss …

As we have started to see, the density of points in PC space may be uniform or not, depending on the data. If the datapoints have been successfully spread out by the PCA, even if a large number of components is required, clustering may succeed in identifying groups of related points.

A very important point is that PCA is a useful tool in working with CNNs. Particularly if one has unlabeled data (unsupervised learning), feature extraction requires interpretation; also, the dimensionality is fixed if using transfer learning (previously developed network). A common approach is to use PCA on the feature vectors to discover which features are most useful, i.e., how few features can explain the variance. Then one can apply cluster analysis, e.g., t-SNE to find out whether there are natural groupings of datapoints.
Discuss …

CCA

- We've seen before that CCA is a close cousin to PCA. Nevertheless, it's a lot more complicated because of having multiple variables (features) on both sides of the input/output divide.
- With samples of materials, it is very common to have multiple sources of variation. Composition is an obvious variable with multiple elements (metals, ceramics), or molecular structure (polymers). The second source of variation is processing applied to the samples such as deformation and heating. All this leads to multiple columns in our spreadsheet (e.g., the HEA database).
- Researching the properties of materials for a specific application means optimizing a combination of properties, not just a single property. Although properties can be combined together in a single figure-of-merit, it is better to consider all the interactions. CCA is a useful tool for this purpose.
  Discuss what materials problems one could apply CCA to, and strengths & weaknesses of the technique …

As with PCA, CCA is useful for analyzing trends from neural nets. As we have seen, a neural net can be configured to output multiple results such as regression and classification combined.

PCA vs. CCA

- PCA is closer to being an exploratory technique in that it makes no assumptions about the significance of the features or variables. It simply tries to diagonalize the data, i.e., inform us where the greatest variance exists and which (combinations of) features contribute most strongly to that variance.
- An example of this application of PCA was given in the NEU defect database analysis. The feature extraction results in rather long feature vectors (one per image) and there is no *a priori* theory that directs us to a particular choice of features to represent each image. Therefore, it is reasonable to apply PCA, not just to reduce the dimensionality but also to trim the number of features. Then the result of that analysis can be pipelined into t-SNE to explore whether a natural clustering exists and whether those clusters correspond to the human-determined labels (as to which defect type is which). In this particular case, there was a trade-off between the cost of the t-SNE calculation (order ~n^2) and the fraction of variance explained by the chosen number of components.
- CCA, by contrast, is more directed. In many cases, we know the phenomenon to be explained and there are various aspects of the processing (e.g.) that were varied. Although this does not dictate the model to be applied, it does focus the search on finding a relationship between two sets of variables (features). In the case of "Parsing abnormal grain growth", Lawrence et al., *Acta Mater*. **103** 681-687 (2016), the phenomenon of abnormal grain growth was obvious but it was not obvious what the best numerical descriptor of the result might be. Moreover, the results pointed to a dependence on composition in the alumina that became semi-obvious in retrospect, i.e., the compositional sensitivity had already been published albeit a long time ago.

Random Forest

- Random Forest is a powerful technique that is robust against many of the typical problems with data. It makes no assumptions about linear relationships, nor about how the data are distributed.
- It is, however, more like an advanced MLR, i.e., pick an output of interest and find out what it depends on.
- The various randomization features seem to be effective at avoiding problems with correlated inputs. The latter issue is very noticeable with PCA and CCA because it leads to a singular matrix, i.e., no solution is available.
- Just as with MLR, determining how strongly any individual variable affects the output requires extra work.
- Discuss the pros & cons of RF vs. variants of MLR …

Neural Nets

- We now understand how neural nets work.
- They are particularly effective at allowing multiple features to influence each other and combine in different ways.
- The variety of activation functions allow for a non-linear response, which turns out to be very important in many problems.
- Discuss how to choose, say, application of an artificial neural net over random forest. Is there, for example, any difference in interpretability?

Convolutional neural nets (CNNs) use the same tools but are more complex so as to be able to use images as input. The typical 224x224 image used as input to VGG16, e.g., means that we literally have 50,176 input variables, which is far too many for a regular NN.

Discuss whether CNNs can be usefully applied to data other than images …

Genetic Programming, Symbolic Regression

- Coming up next week is symbolic regression.
- The idea is to use machine learning to discover mathematical relationships between variables.
- In other words, we expect a definite relationship between the output of interest and the independent variables but we do not know *a priori* what the functions might be.
- This is then an explicit way to tackle non-linearity in our data.
- The method that we shall use makes use of genetic programming in which multiple models are evaluated in parallel and allowed to compete and give rise to new combinations of models.

Time Series

- An entire additional field of application of machine learning is to time series.
- Although most people think of this as applying to acoustic data, it is generally applicable to any problem in which there is a definite sequence of datapoints. This can be images, weather data, simulations of deformation etc.
- E.g., https://aihubprojects.com/weather-prediction-using-ml-algorithms-ai-projects/
- A classical analysis method is to apply a Fourier transform to the data.
  Discuss what information this provides: does it show long-term trends? If not, what might one use instead?
- More sophisticated methods:
  e.g., ARIMA, which stands for Auto-Regressive Integrated Moving Average.
- In general, we typically assume that any time signal can be split up: *Trend* refers to the gradual increase or decrease of the time series over time, *Seasonality* is a repeating short-term cycle in the series, and *Error* refers to noise not explained by seasonality or trend. Moving average therefore refers to the number of lags of the error component to consider.
- Interesting example from materials recycling: Big data for time series and trend analysis of polymer waste management in India, Materials Today: Proceedings, Volume 37, Part 2, 2021, Pages 2607-2611. https://doi.org/10.1016/j.matpr.2020.08.507

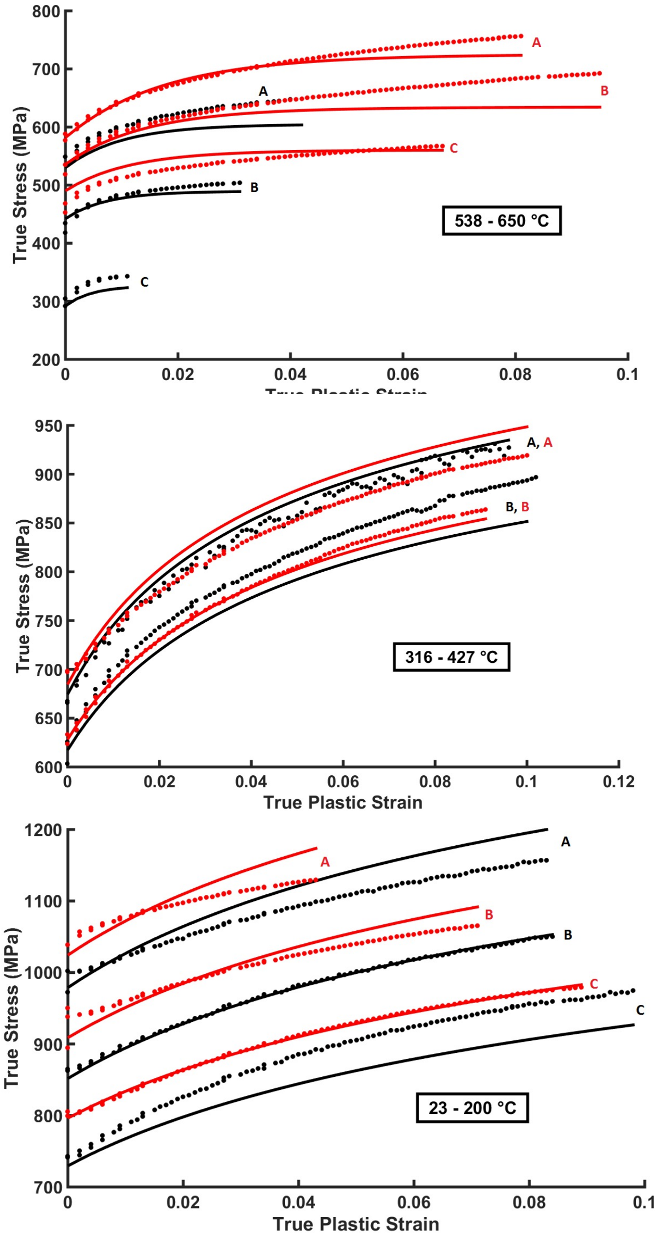
Data

- We have discussed the availability of data. In general, much of the existing data pertains to atomic-scale simulations of materials. Data on structural materials is limited and mostly found in datasheets provided by companies that supply feedstock, e.g., for additive manufacturing.
- FAIR Data Principles have been established. These principles have been adopted by several major players in global data management (see Table 1). The ideas include making data Findable, Accessible, Interoperable (flexible), and Reusable.
- This article provides an overview of what is currently available: *Adv. Sci*. 2019, **6**, 1900808, DOI: 10.1002/advs.201900808. Also uploaded to Canvas.

Multivariate Analysis of Constitutive Relations

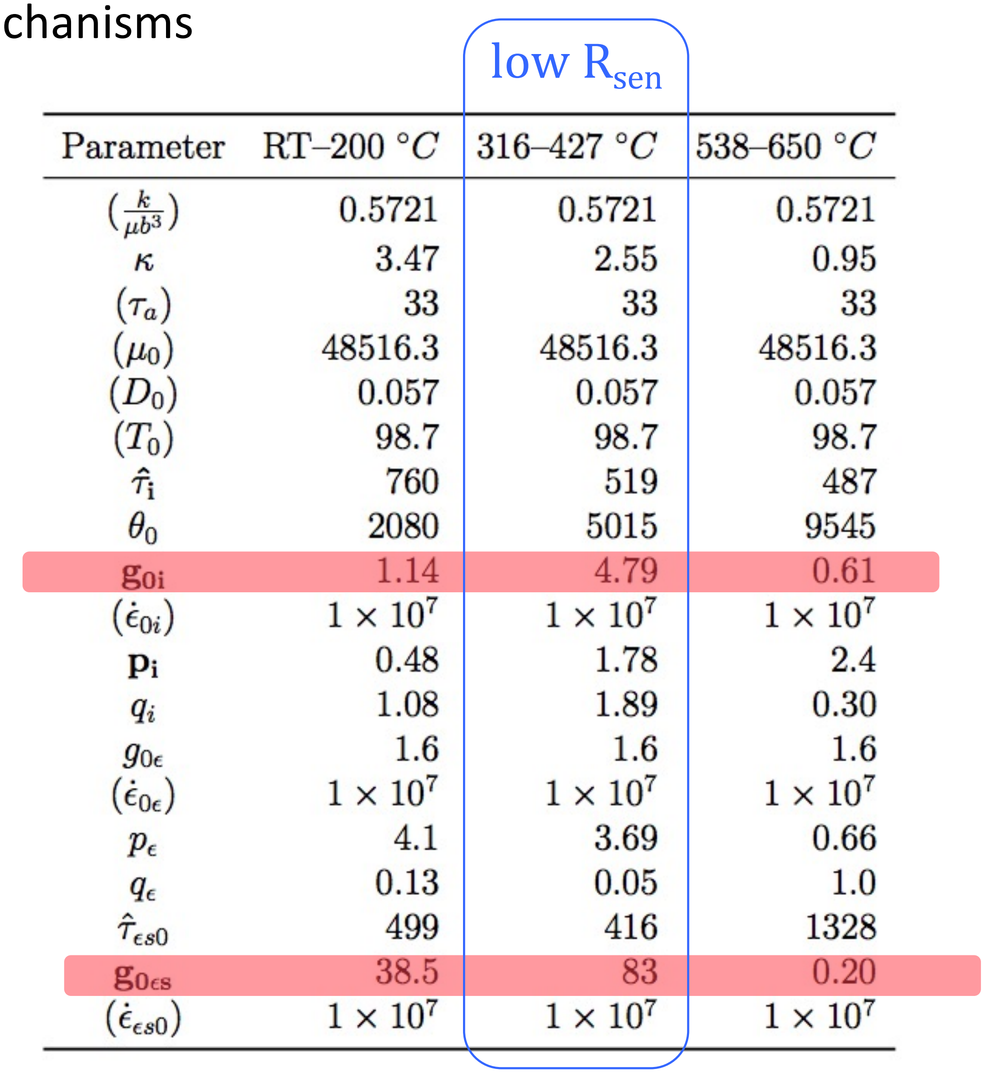
- A particular strength of the methods discussed in class is the ability to analyze lots of variables all at once.
- “Application of **canonical correlation analysis** to a sensitivity study of constitutive model parameter fitting” S. Mandal, B.T. Gockel, and A.D. Rollett, *Materials and Design* **132**, 30-43 (2017) represents an example of sensitivity analysis. A complex model of mechanical strength and its evolution with strain, strain rate and temperature was analyzed. The (Mechanical Threshold Stress *aka* MTS) model has 19 variables in its basic form. Varying each of these one at a time is expensive and prone to miss interactions. The re-statement of this is that binary correlations between multiple measures of the output of the model and the model parameter values would miss synergistic effects.
- One very important aspect of this work was indirect validation through analysis of experimental results. When parameters were developed for the MTS model by fitting to data for the stress-strain behavior of Ti6242, which is an alloy closely related to Ti-6Al-4V, as a function of temperature and strain rate, it turned out that there was a “flat spot” at intermediate temperatures. By flat spot, we mean that there was little variation in flow stress caused by changes in temperature or strain rate, which is labeled “athermal behavior”. This, in the MTS fit, was associated with a large barrier to dislocation motion. Given that the signature of thermally assisted dislocation motion is strong sensitivity to temperature (and mild sensitivity to strain rate), this outcome was in good agreement with the results of the sensitivity study.
- The work particularly highlighted the importance of quantifying relationships between variables on a multi-variate basis and assuming that the inputs/variables/features are *not* independent.

Validation of CCA Predictions: Experiment



Exp: points, MTS: line; Black: Quasi-static strain rate, Red: High strain rate (0.01/s)

- Uniaxial compression of Ti-6242 (**Courtesy: Dr. Brian Gockel, AFRL**)
- MTS model fitted separately for the three different temperature regimes with varying deformation mechanisms



low R_{sen}

Parameter	RT–200 $°C$	316–427 $°C$	538–650 $°C$
$(\frac{k}{\mu b^3})$	0.5721	0.5721	0.5721
κ	3.47	2.55	0.95
(τ_a)	33	33	33
(μ_0)	48516.3	48516.3	48516.3
(D_0)	0.057	0.057	0.057
(T_0)	98.7	98.7	98.7
$\hat{\tau}_{\mathrm{i}}$	760	519	487
θ_0	2080	5015	9545
$\mathbf{g_{0i}}$	1.14	4.79	0.61
$(\dot{\epsilon}_{0i})$	1×10^7	1×10^7	1×10^7
$\mathbf{p_i}$	0.48	1.78	2.4
q_i	1.08	1.89	0.30
$g_{0\epsilon}$	1.6	1.6	1.6
$(\dot{\epsilon}_{0\epsilon})$	1×10^7	1×10^7	1×10^7
p_ϵ	4.1	3.69	0.66
q_ϵ	0.13	0.05	1.0
$\hat{\tau}_{\epsilon s0}$	499	416	1328
$\mathbf{g_{0\epsilon s}}$	38.5	83	0.20
$(\dot{\epsilon}_{\epsilon s0})$	1×10^7	1×10^7	1×10^7

Discussion

- CCA was used in the example above.
- Should other techniques have been tried? If so, why?

Example from OPV research

"Despite the recent fast progress in materials informatics and data science, data-driven molecular design of organic photovoltaic (OPV) materials remains challenging. We report a screening of conjugated molecules for polymer−fullerene OPV applications by supervised learning methods (artificial neural network (ANN) and random forest (RF)).

We report a screening of conjugated molecules for polymer−fullerene OPV applications by supervised learning methods (artificial neural network (ANN) and random forest (RF)). Approximately 1000 experimental parameters including power conversion efficiency (PCE), molecular weight, and electronic properties are manually collected from the literature and subjected to machine learning with digitized chemical structures. Contrary to the low correlation coefficient in ANN, RF yields an acceptable accuracy, which is twice that of random classification."

Results based on 1200 points from 500 papers.

Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest, S. Nagasawa et al, *J. Phys. Chem Lett.* **9**, 2639 (2018)

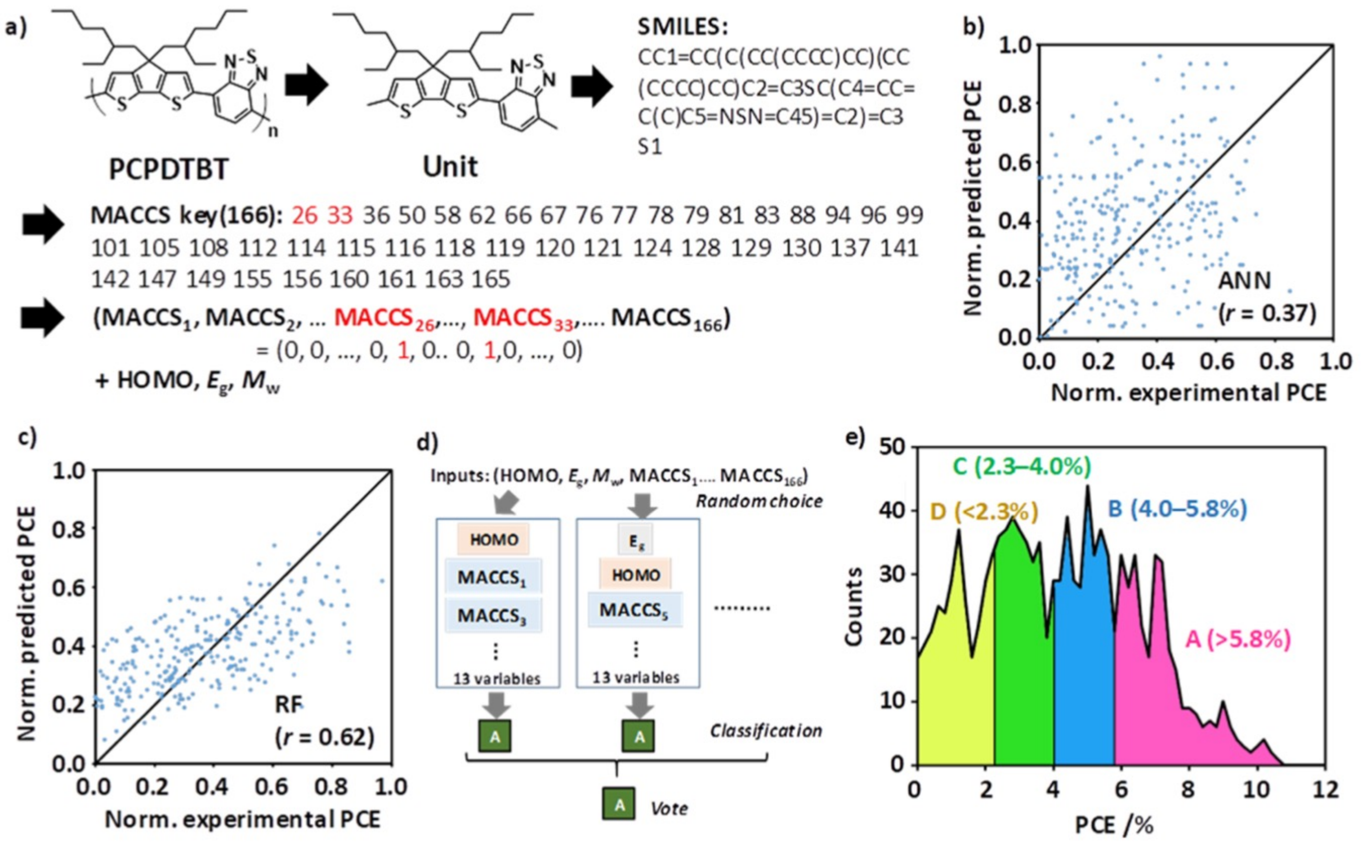


Figure 2. (a) Scheme of converting a chemical structure to digitized data. (b) Results of ANN and (c) results of RF, where the horizontal and vertical axes represent the normalized experimental PCE and predicted PCE, respectively, and r is the correlation coefficient. The diagonal black line indicates the perfect positive correlation (r = 1). (d) Scheme of a classification using RF. (e) Histogram of the collected experimental PCE data (~1200) and classification of n = 4. For example, label "A" corresponds to the highest PCE group.

Artificial Neural Nets (ANN) led to a relation with r=0.37, which is not acceptable.

They represented PCE in 4 groups (e) and used the RF in (d).

Based in part on the RF results, they demonstrated an alternative approach to the design of polymers for OPVs.

Example from OPV research

Supervised vs. Unsupervised Learning

- Supervised learning comprises much of what we use machine learning for.
- Among the various problems with this basic approach, it assumes that data comes with labels. We in MSE often do not consider this question because in our often small (tiny?) data sets, labeling is often built-in to the results. Another issue in big data is the cost of labeling. This activity employs a large workforce these days (e.g., someone has to look at any given photograph and decide what kinds of objects it shows).
- Even if we have labels, we do not necessarily know that they are correct or even reasonable. For example, in the HEA dataset, we asked you to apply the group IDs as labels to each alloy and then use the clustering analysis to find out whether the data analytics would result in the same set of groups or something quite different.
- If the result of data analytics points to a different model than the initial hypothesis, this is very important (and perhaps worthy of publication). It at least required that statistical tests of significance of competing models be conducted. This was examined in the context of CCA.

Too Many Variables!

- Data being expensive to acquire, it can often happen that we have too few data points and too many variables/features.
- In simple terms, the spreadsheet has many more columns than rows.
- We dealt with this in the discussion of the LASSO method. To some approximation, we use an algorithm that tries MLR with as many combinations as possible of N variables where N is varied from 1 up to some limit (perhaps determined by the number of datapoints).
- As always, one must partition the data, keep back a subset for validation, determine a model with, e.g., LASSO or FeaLect and then test the result. See the work of Kantzos et al.

Summary

- Choosing a reasonable approach should be guided by a cascade of decisions.
- Do you know how to extract information from your data?
  If No, can you make up derived/transformed variables that might allow correlations to be identified? Or, do you need computer vision techniques to extract information from images? Or, do you need time-series analysis for sequences of similar but varying signals/images?
- Do you have a pre-conceived model for your data?
  If No, then use exploratory techniques first
  If Yes, then test multiple models and remember to apply statistical significance tests