

Machine Learning

vs

Materials Informatics

Materials Informatics differs substantially from traditional machine learning

Aspect	Traditional Machine Learning	Materials Informatics
Data Availability (number of data records available)		
Task		???
Data heterogeneity		???
Uncertainty quantification		
Features		
Interpretability		

Materials Informatics datasets are typically pretty small compared to traditional ML

Skip to content

≡ kaggle

Search

Sign In

Register

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset



Search datasets

Filters

Datasets Tasks Computer Science Education Classification Computer Vision NLP Data Visualization

Trending Datasets

See All



Restaurant Reviews · [View](#) · [Edit](#) · [Share](#)

lucif3r · Updated 2 hours ago
Usability 8.8 · 24 KB
1 Task · 1 File (other)

1



Market Basket Optimisation · [View](#) · [Edit](#) · [Share](#)

lucif3r · Updated 2 hours ago
Usability 8.2 · 47 KB
1 Task · 1 File (CSV)

0



Flower Photos · [View](#) · [Edit](#) · [Share](#)

Batool Abbas · Updated 2 hours ago
Usability 8.8 · 219 MB
3671 Files (other)

0

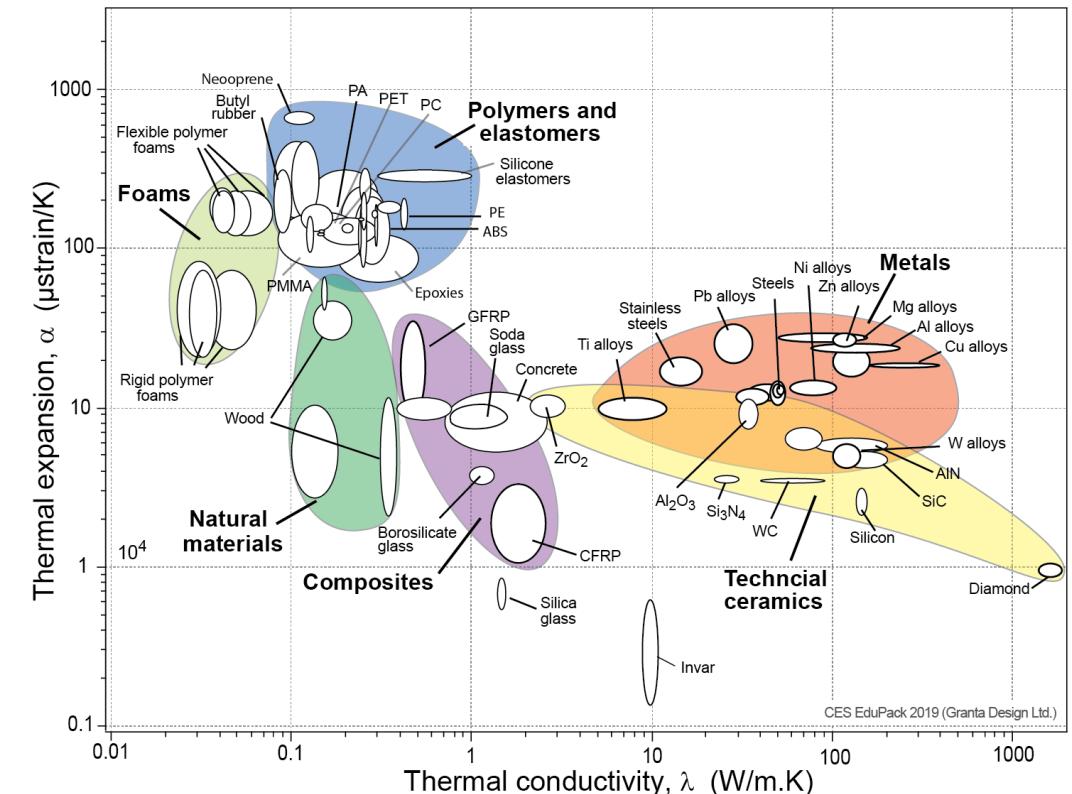
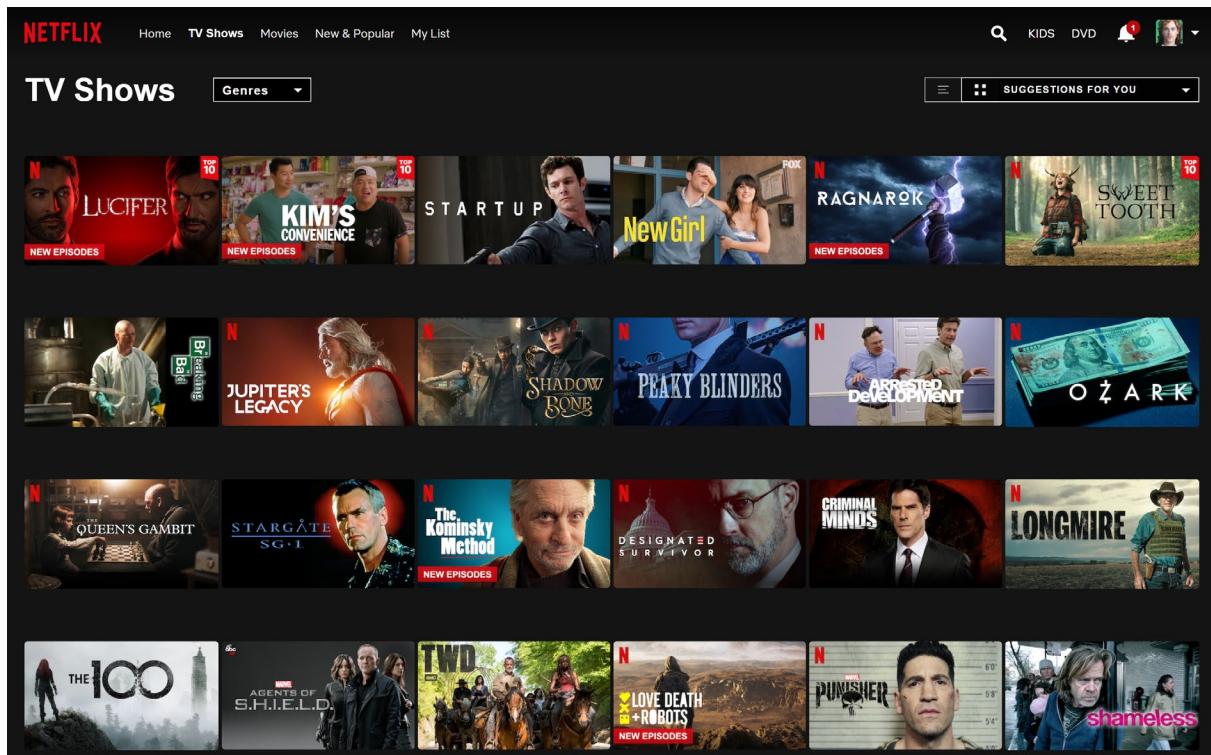


Social Network Ads · [View](#) · [Edit](#) · [Share](#)

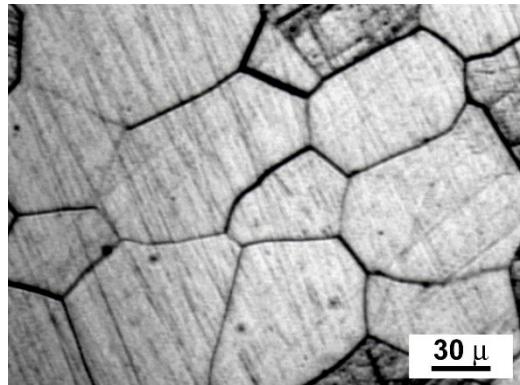
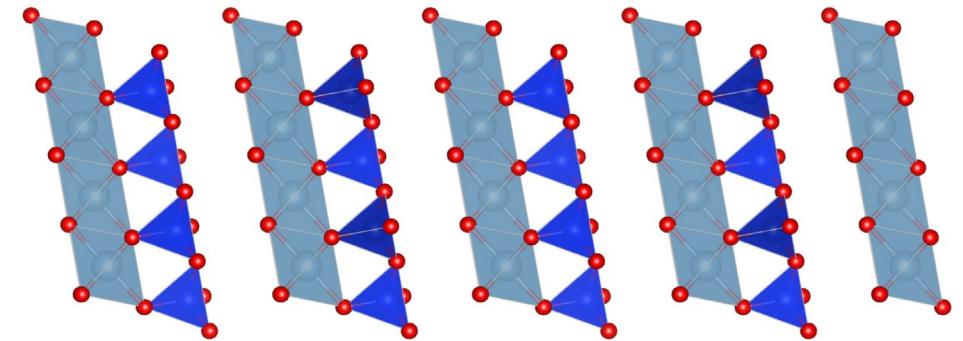
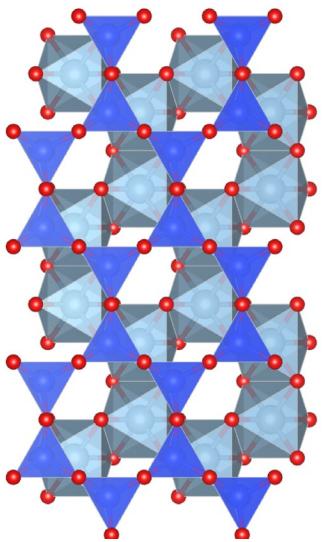
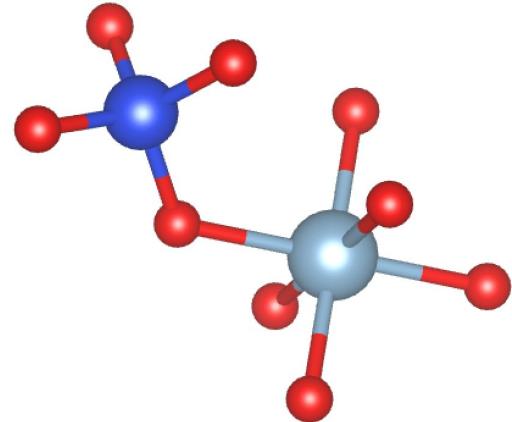
lucif3r · Updated 5 hours ago
Usability 9.4 · 1 KB
1 Task · 1 File (CSV)

5

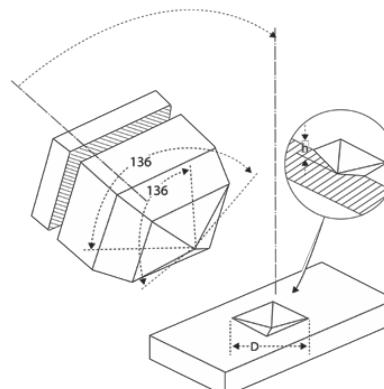
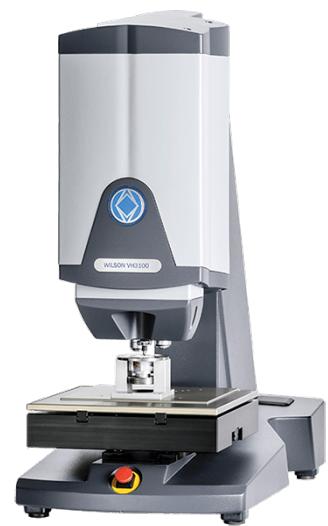
Materials Informatics usually interests itself in identifying unusual outliers rather than averages



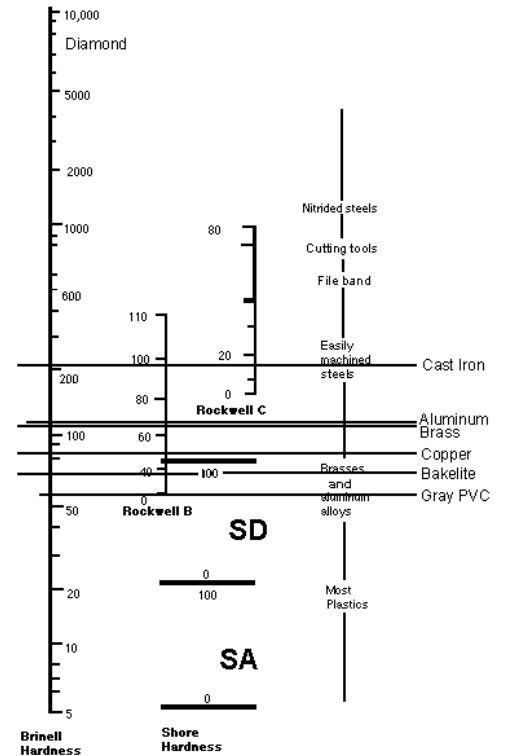
Materials research suffers from heterogeneous data with many modalities



Materials research suffers from heterogeneous data with many modalities



Aproximate Comparison of Hardness Scales



Uncertainty quantification is extremely important when each new data is expensive

Facilities and physical lab space required \$\$

Materials need to be purchased \$\$

Uncertainty quantification is extremely important when each new data is expensive

Samples need to be synthesized 

High chance of failure 

Diverse reasons for failure 

Uncertainty quantification is extremely important when each new data is expensive

Samples need to be characterized \$\$ & ⏳ ⏳

Characterization requires equipment \$\$

Characterization requires expertise \$\$

Characterization takes time ⏳ ⏳

Compositional design space is enormous!

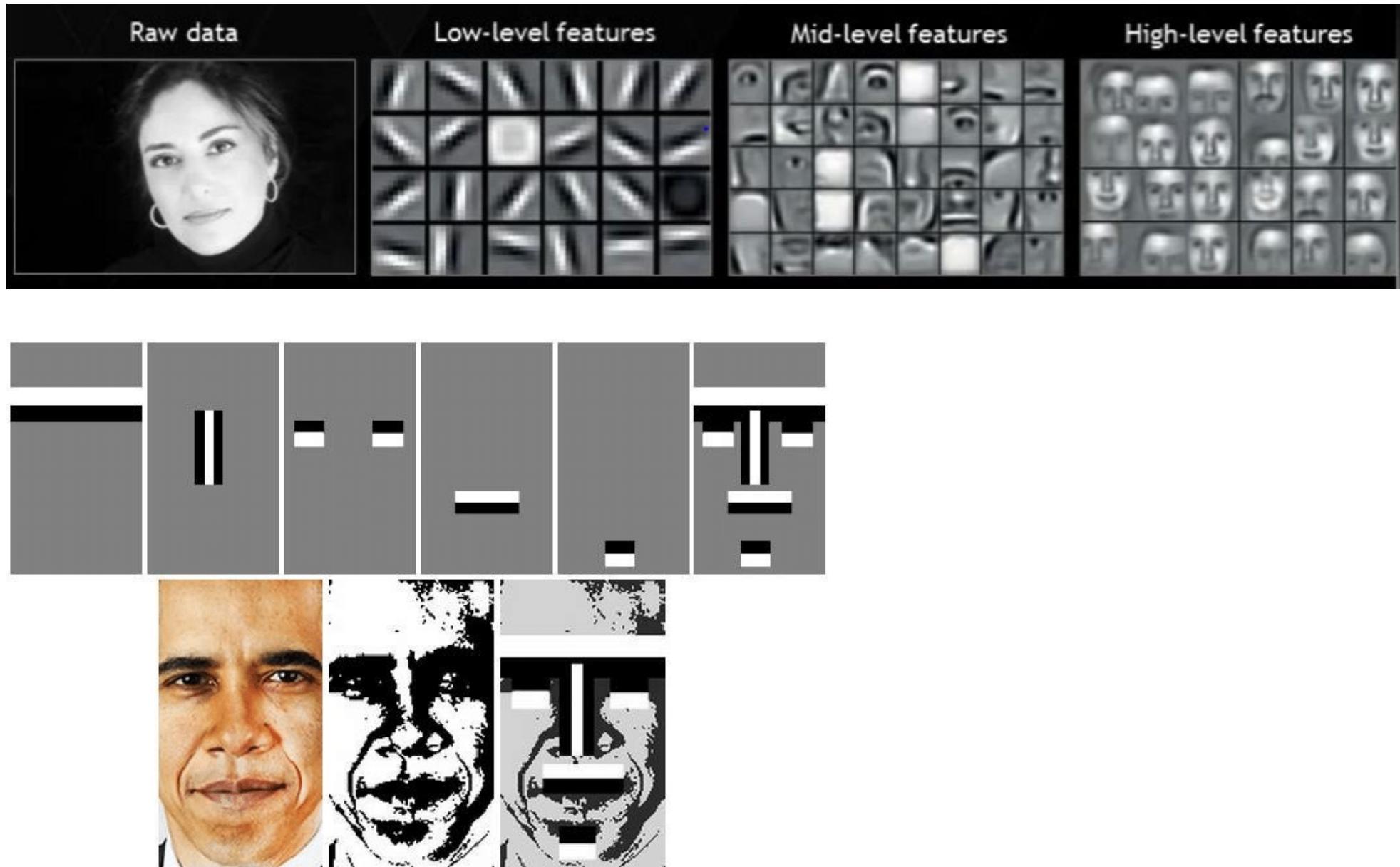
Total unique inorganic compounds $\approx 10^{12}$

Four component systems
 $(A_aB_bC_cD_d)$.

Ignore dopants
 $(a, b, c, \text{ and } d > 0.03)$.

<https://htwins.net/scale2/>

Traditional ML has moved away from feature engineering to deep learning



Feature engineering can significantly improve predictions with $<10^3$ data instances



Technical Article | Published: 27 August 2020

Is Domain Knowledge Necessary for Machine Learning Materials Properties?

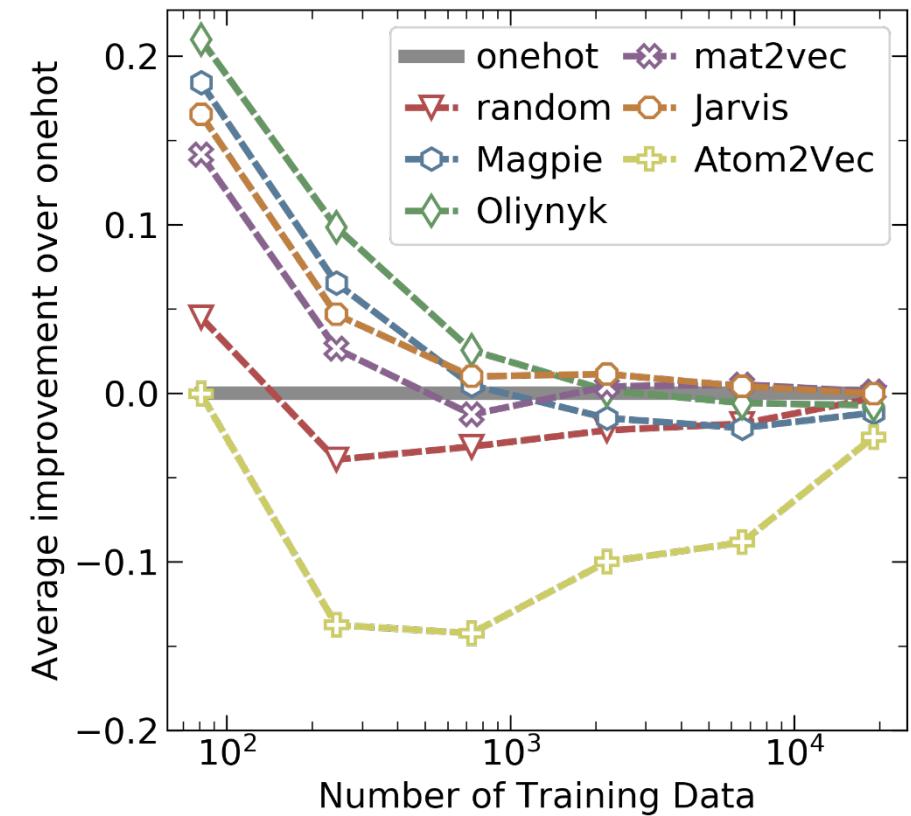
Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang & Taylor D. Sparks

Integrating Materials and Manufacturing Innovation 9, 221–227 (2020) | [Cite this article](#)

468 Accesses | 1 Altmetric | [Metrics](#)

Abstract

New featurization schemes for describing materials as composition vectors in order to predict their properties using machine learning are common in the field of Materials Informatics. However, little is known about the comparative efficacy of these methods. This work sets out to make clear which featurization methods should be used across various circumstances. Our findings include, surprisingly, that simple fractional and random-noise representations of elements can be as effective as traditional and new descriptors when using large amounts of data. However, in the absence of large datasets or for data that is not fully representative, we show that the integration of domain knowledge offers advantages in predictive ability.



Feature engineering can significantly improve predictions with $<10^3$ data instances



Technical Article | Published: 27 August 2020

Is Domain Knowledge Necessary for Machine Learning Materials Properties?

Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang & Taylor D. Sparks

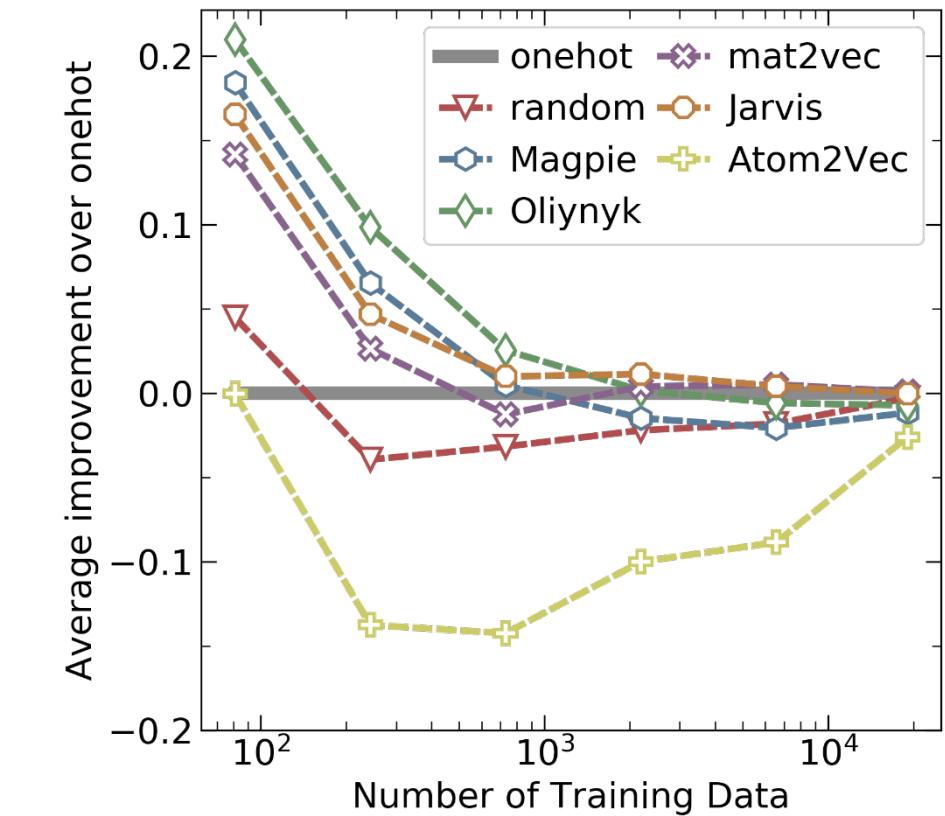
Integrating Materials and Manufacturing Innovation 9, 221–227 (2020) | [Cite this article](#)

468 Accesses | 1 Altmetric | [Metrics](#)

Abstract

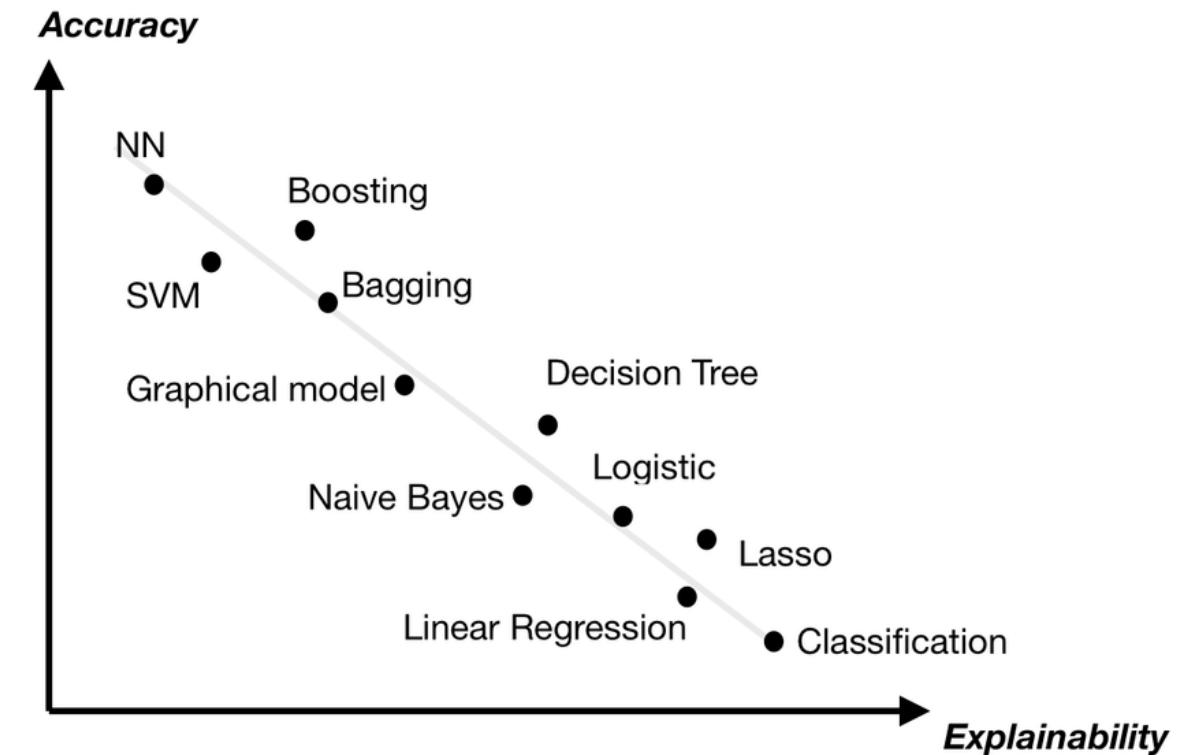
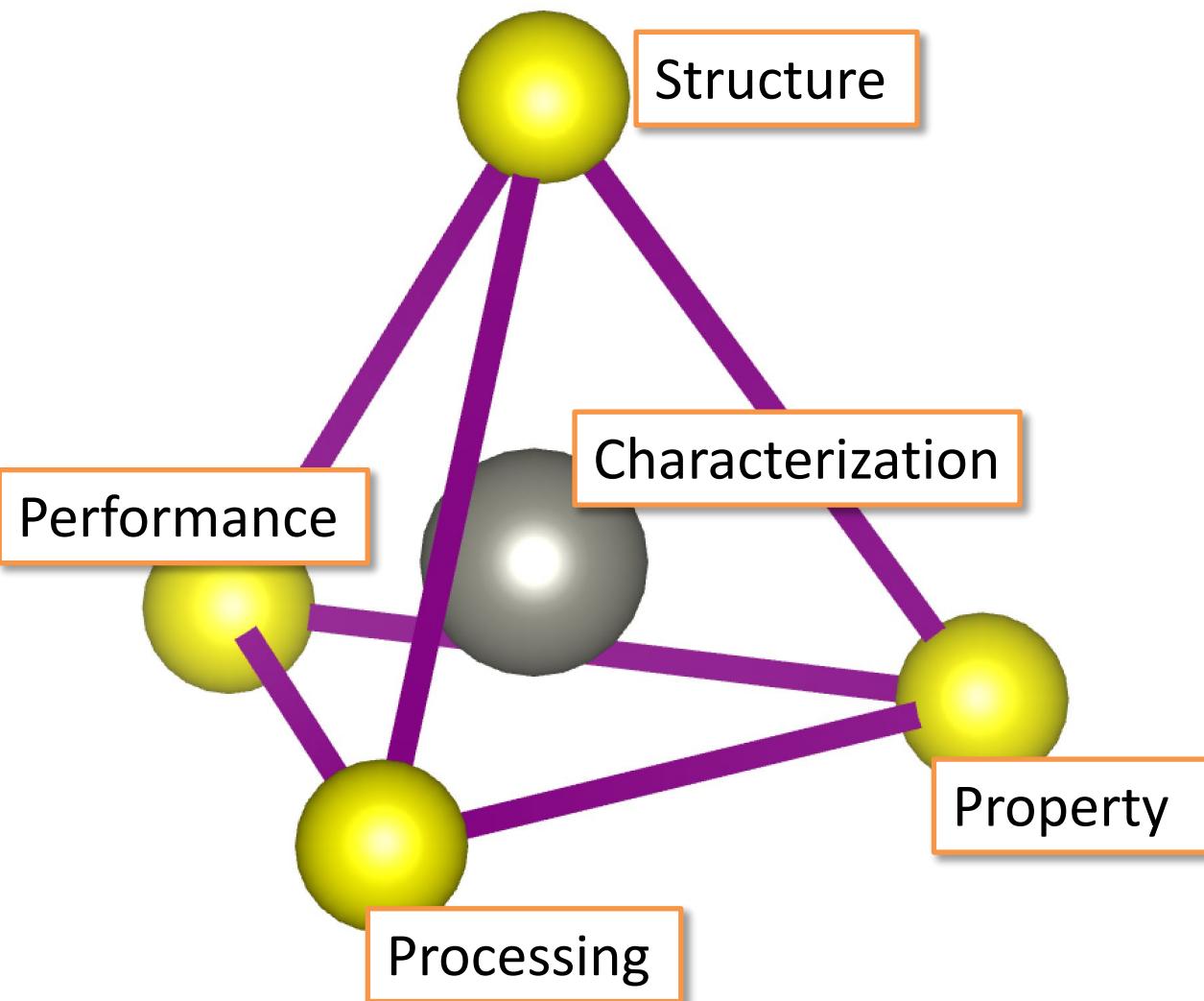
New featurization schemes for describing materials as composition vectors in order to predict their properties using machine learning are common in the field of Materials Informatics.

However, little is known about the comparative efficacy of these methods. This work sets out to make clear which featurization methods should be used across various circumstances. Our findings include, surprisingly, that simple fractional and random-noise representations of elements can be as effective as traditional and new descriptors when using large amounts of data. However, in the absence of large datasets or for data that is not fully representative, we show that the integration of domain knowledge offers advantages in predictive ability.



Anton Oliynyk!

Structure-property-processing linkages depend on interpretable relationships



Materials Informatics differs substantially from traditional machine learning

Aspect	Traditional Machine Learning	Materials Informatics
Data Availability (number of data records available)	Very large, typically $>10^6$ and data <i>down sampling</i> is common for computational tractability	Small, typically $<10^3$
Task	Identify value for given input	Identify extremes (large or small) for given input
Data heterogeneity	Possible to achieve high homogeneity	Typically multimodal with wide heterogeneity
Uncertainty quantification	Not important	Important
Features	Not important with deep learning from large data	Very important due to limited data
Interpretability	Typically less important	Very important

Materials Data: Repositories

