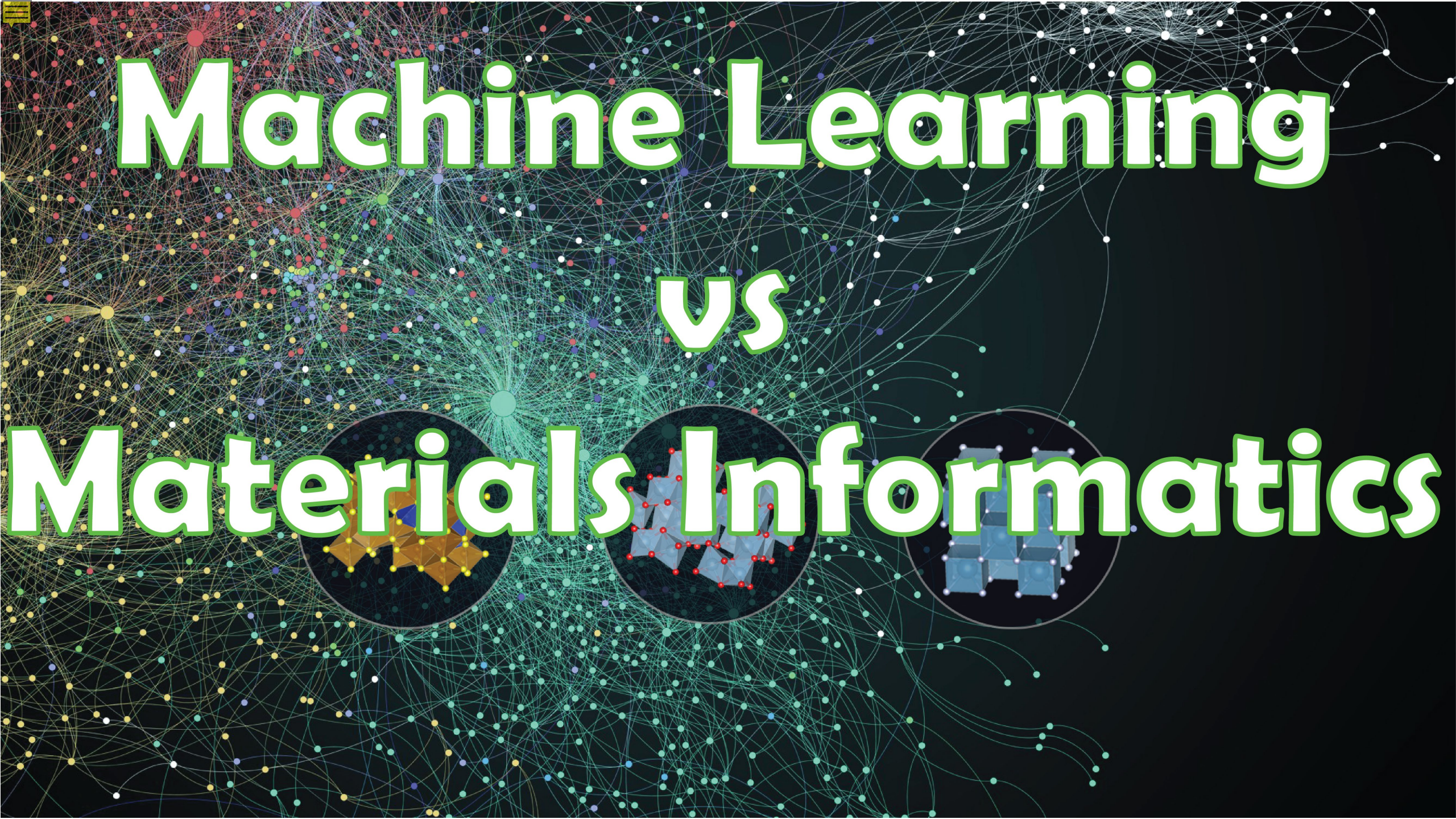
Machine Learning vs Materials Informatics



Machine learning and materials informatics differ substantially!

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 very small n<1000

kaggle

Search

Sign In

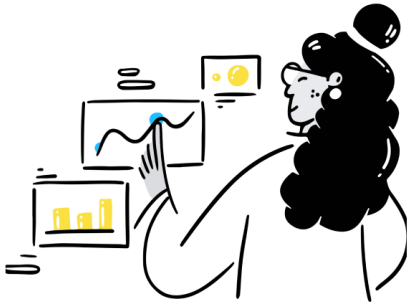
Register

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

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

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

1

Market Basket Optimisation

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

0

Flower Photos

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

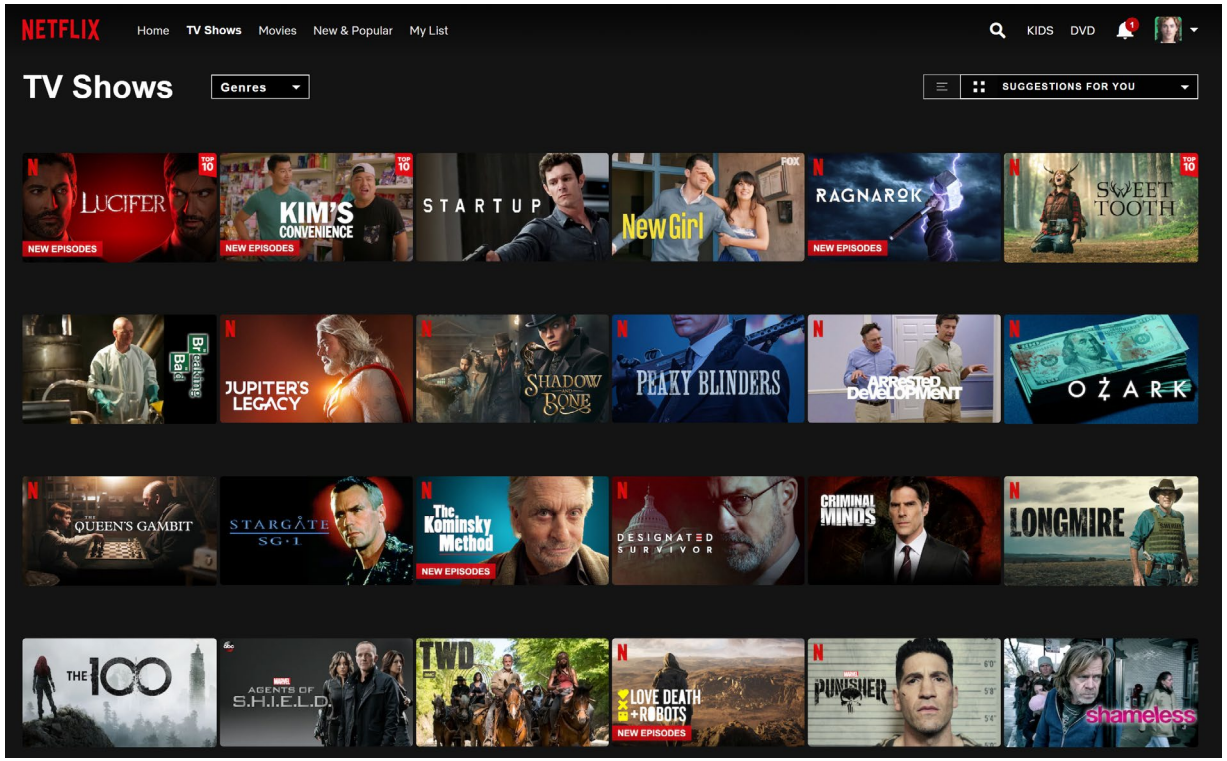
0

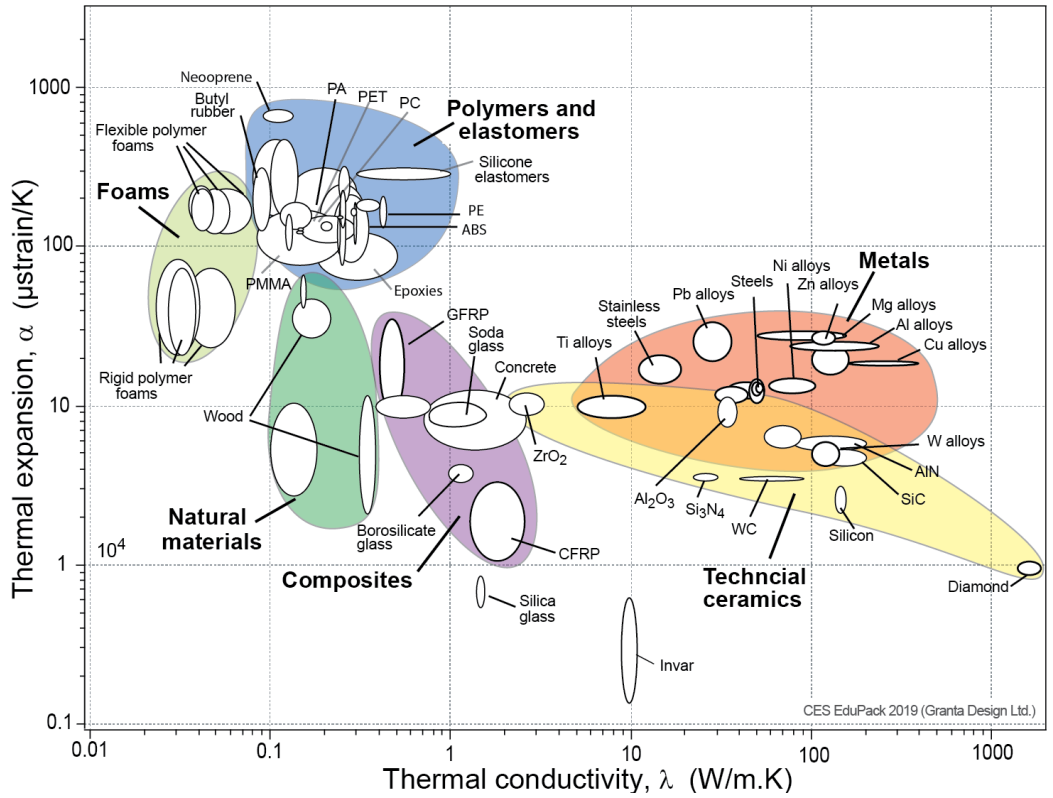
Social Network Ads

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

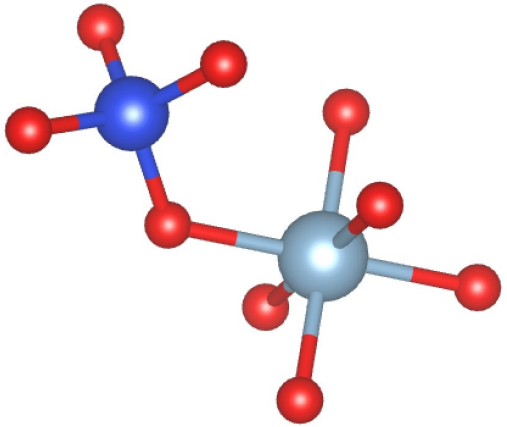
5

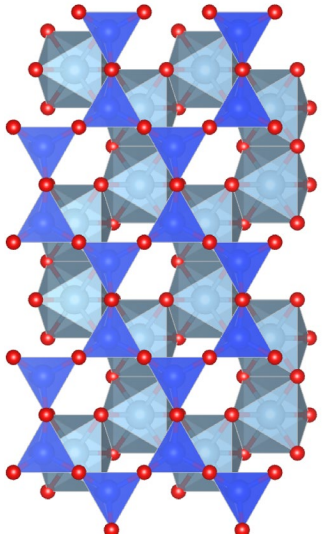
Materials informatics is more interested in outliers and extrapolation

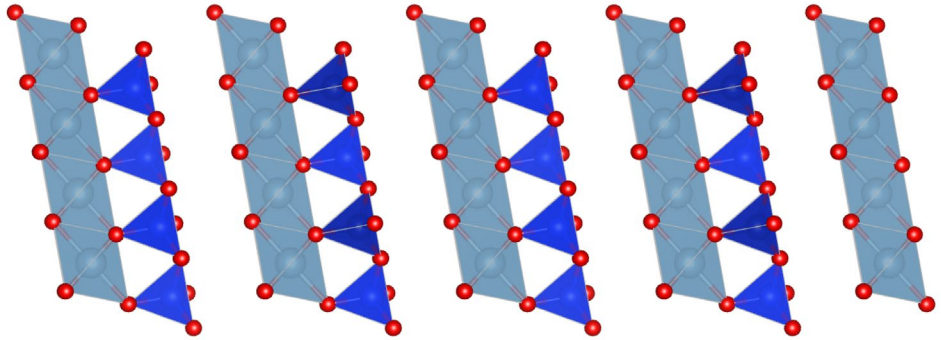


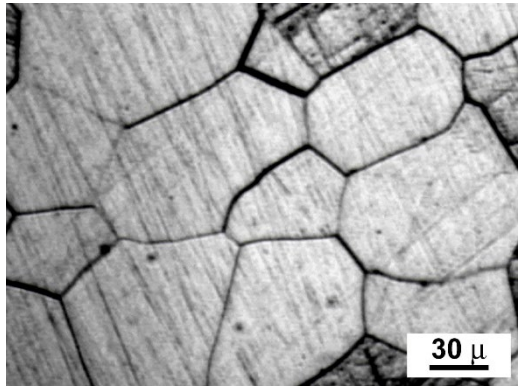


Materials data is heterogeneous and highly multimodal

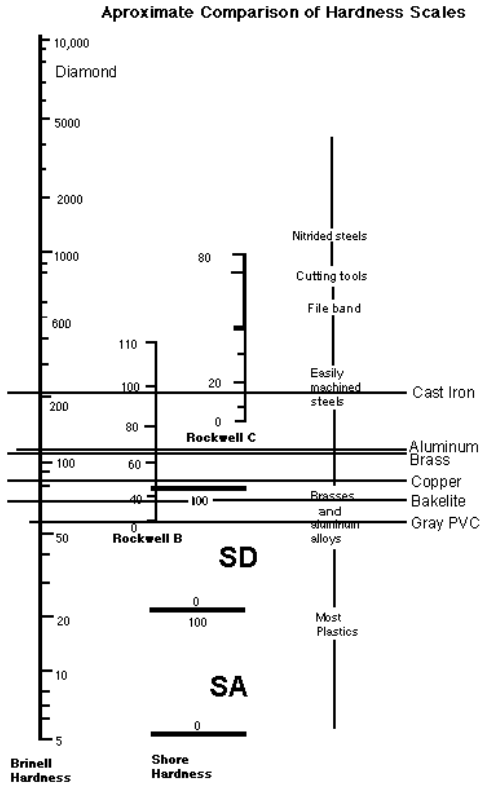


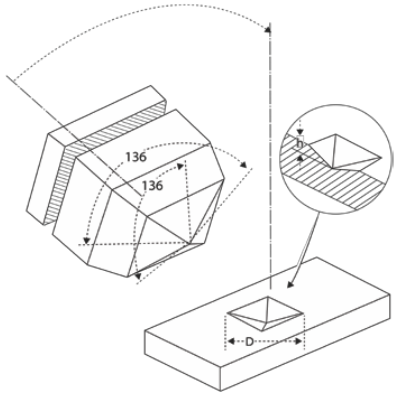






Materials data is heterogeneous and highly multimodal





Uncertainty quantification is important when new data is expensive

Facilities and physical lab space required **$$**

Materials need to be purchased **$$**

Uncertainty quantification is important when new data is expensive

Samples need to be synthesized ⌛⌛

High chance of failure ⌛⌛

Diverse reasons for failure ⌛⌛

Uncertainty quantification is important when new data is expensive

Samples need to be characterized $$ & ⏳⏳

Characterization requires equipment $$

Characterization requires expertise $$

Characterization takes time ⏳⏳

Compositional space is enormous!

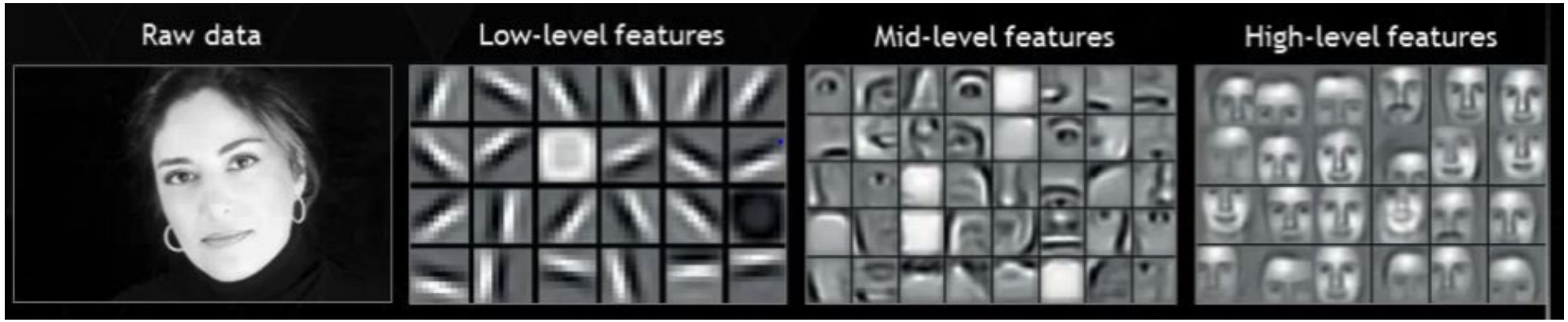
Total unique inorganic compounds ≈ 10^{12}

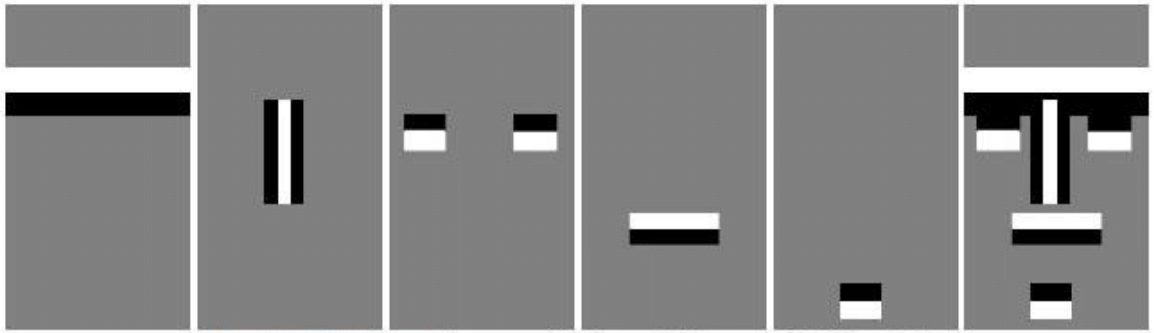
Four component systems
($A_aB_bC_cD_d$).

Ignore dopants
(a, b, c, and d > 0.03).

https://htwins.net/scale2/

Traditional ML has moved away from feature engineering





Feature engineering is still valuable when data n<10^3

Springer Link

Technical Article | Published: 27 August 2020

Is Domain Knowledge Necessary for Machine Learning Materials Properties?

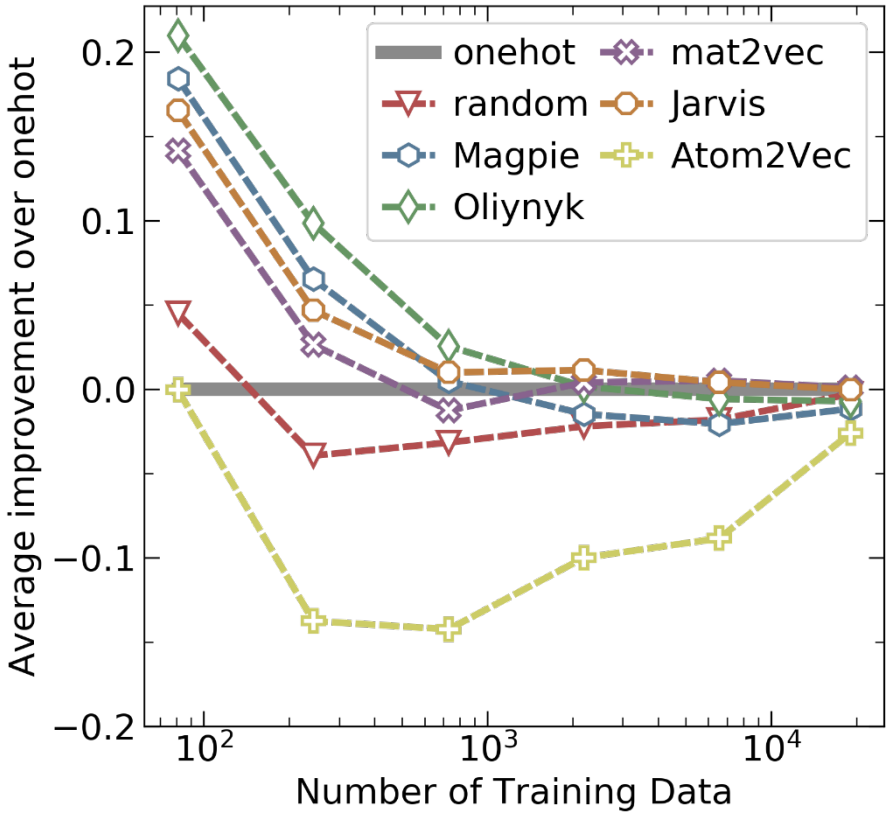
Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang & Taylor D. Sparks

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 is still valuable when data n<10^3

Springer Link

Technical Article | Published: 27 August 2020

Is Domain Knowledge Necessary for Machine Learning Materials Properties?

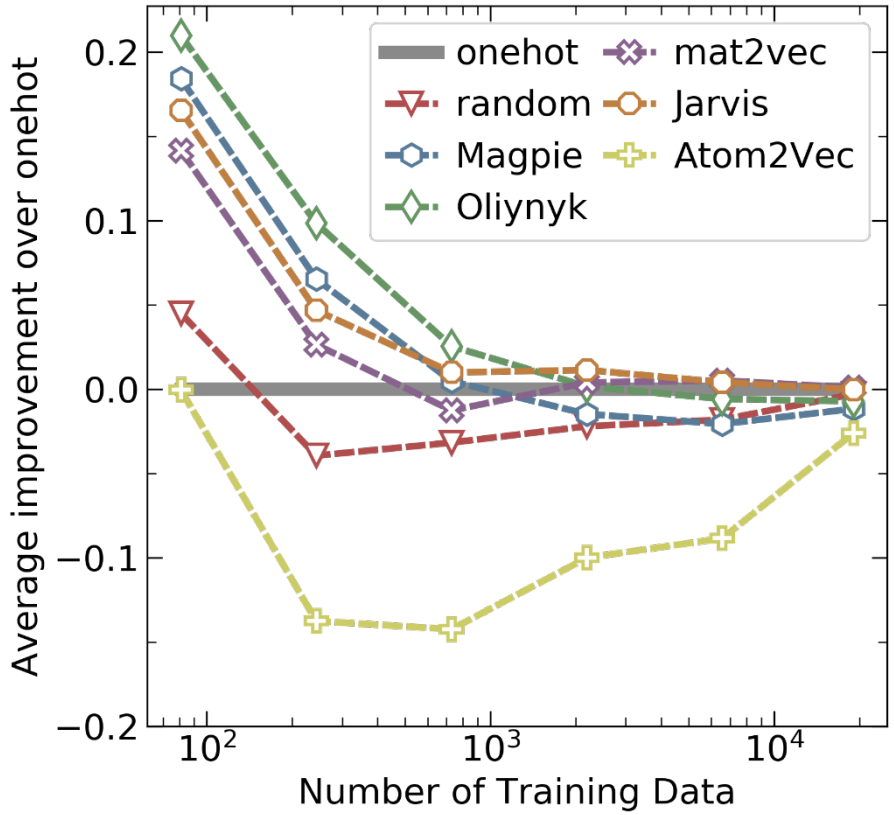
Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang & Taylor D. Sparks

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

468 Accesses | **1** Altmetric | Metrics

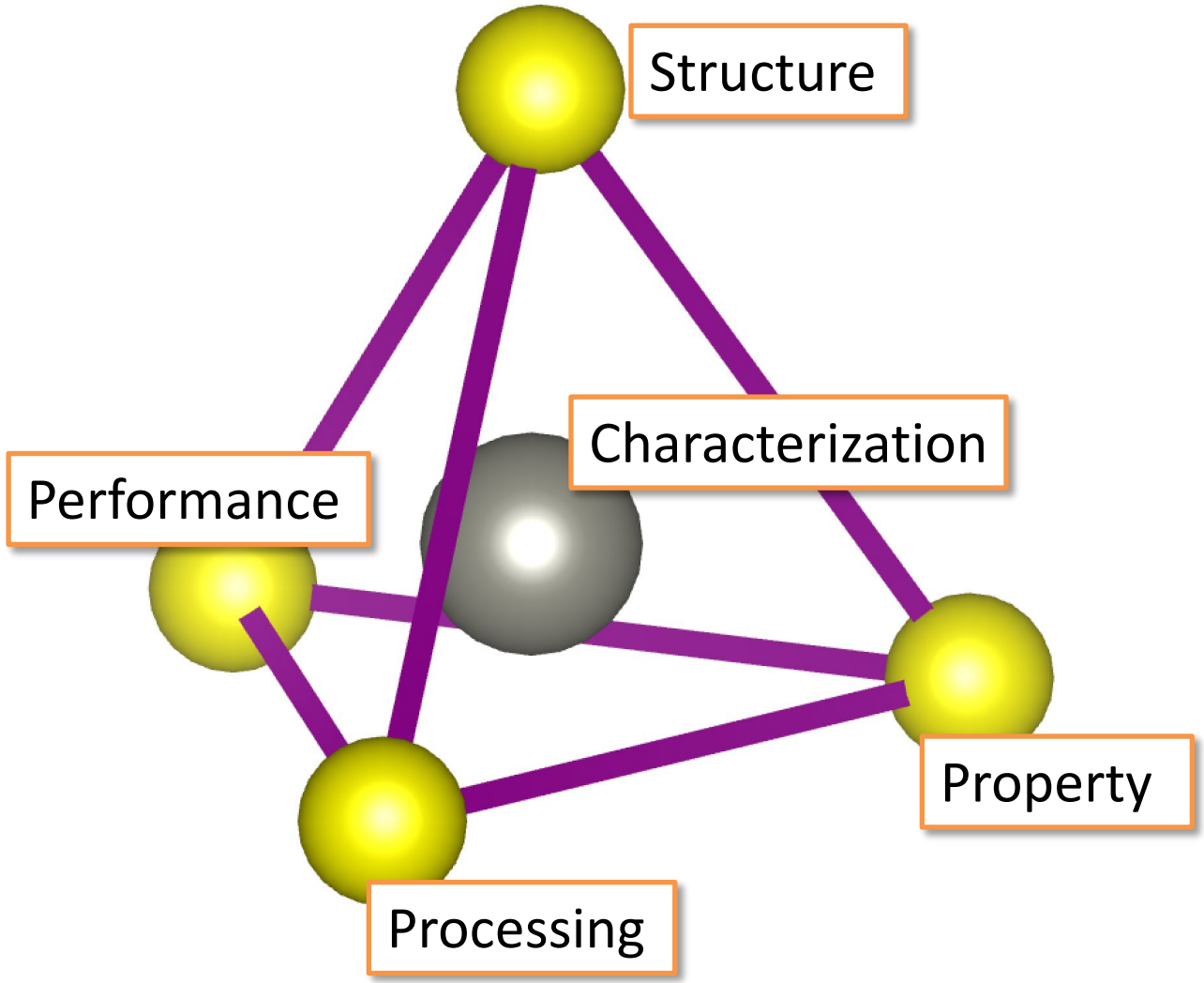
Abstract

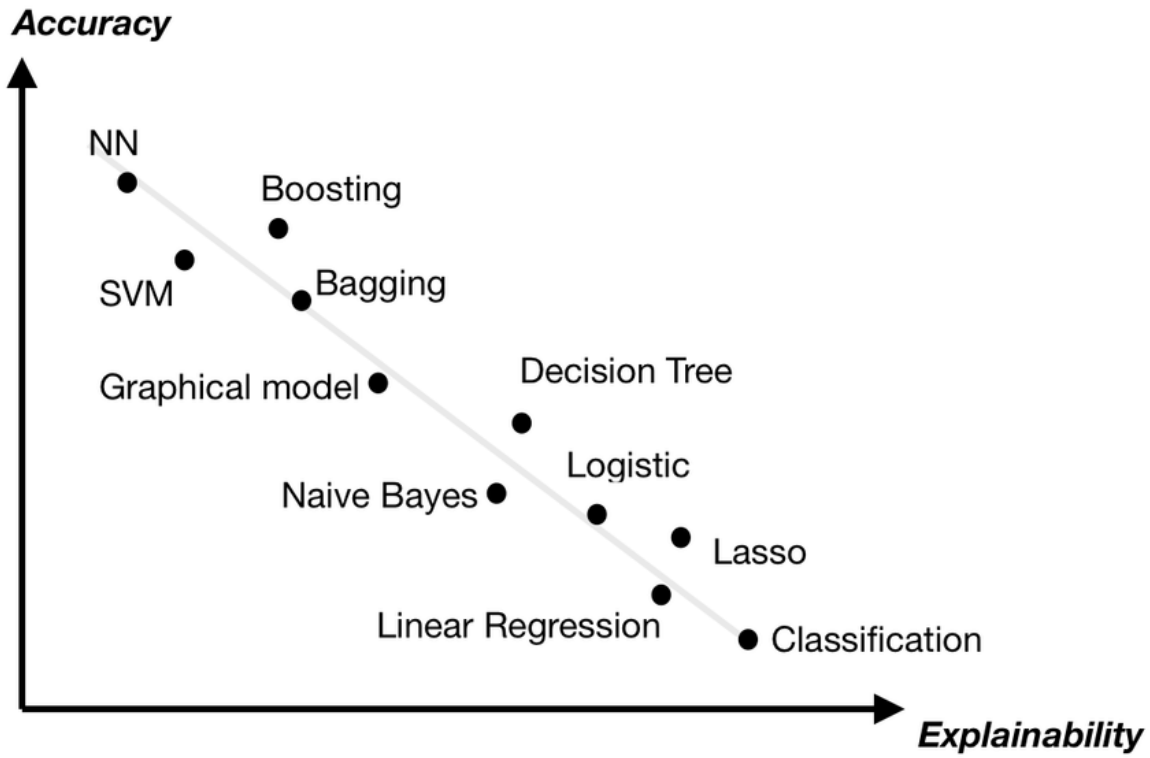
New featurization schemes for describing materials as composition vectors in order to predict their properties using machine learning are common in the field of Materials Informatics. However, little is known about the comparative efficacy of these methods. This work sets out to make clear which featurization 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-Processing-Property linkages depend on interpretability





Machine learning and materials informatics differ substantially!

Aspect	Traditional Machine Learning	Materials Informatics
Data Availability (number of data records available)	Very large, typically $>10^6$ and data *down sampling* 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