Feeding Machine Learning Algorithms with Massive Model Suites

Leo Portes[1,2], Mark Jessell[1,2,3], Mark Lindsay[2,4], Guillaume Pirot[1,2,3], Michel Nzikou[1,3], Ed Cripps[1,2]

[1]UWA, [2]ITTC DARE, [3]MinEx CRC, [4]CSIRO

MinEx CRC

Loop

DARE
ARC Training Centre in
Data Analytics for Resources & Environments

THE UNIVERSITY OF WESTERN AUSTRALIA

Agenda

- Background: Noddy & Guo *et al,* 2021
- High dimensionality: Haralick features
- Exploring separability
- Simpler ML architectures
  - SVM, KNN, RF, NN
- Geological Interpretation (hands-on, 3-4 pm)
- Current and future work
  - Real-world data
  - NoddyVerse with 343 classes and 310K data points (hands-on, 3-4 pm)
  - “Maths”+Theory

Background

Noddy & Guo *et al* 2021

Computers & Geosciences 149 (2021) 104701

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

3D geological structure inversion from Noddy-generated magnetic data using deep learning methods

Jiateng Guo [a,*], Yunqiang Li [a], Mark Walter Jessell [b], Jeremie Giraud [b], Chaoling Li [c], Lixin Wu [d], Fengdan Li [c], Shanjun Liu [a]

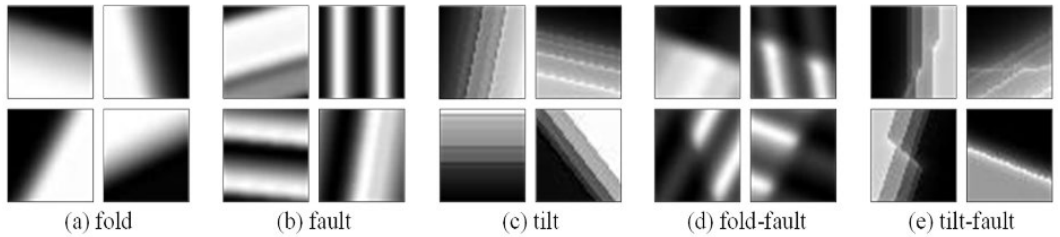


Fig. 1. Samples of the magnetic dataset generated by Noddy. These images are normalized before training.

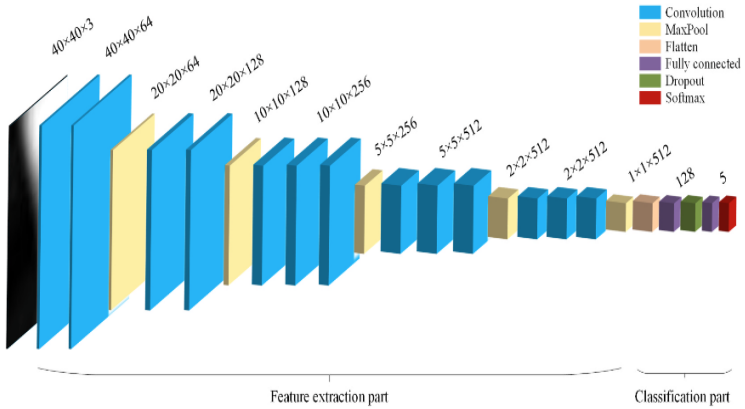


Fig. 3. Network structure of CNN classification model. The feature extraction layer here takes VGG16 as an example. When training with Inception-v3 and ResNet50, the corresponding feature extraction layer is replaced accordingly.

Guo et al, 2021

Goal: predict the parameters of the geological structure:

"a" - dip direction

"b" - dip angle of the fold axial surface

Magnetic field 2D images

Noddy dataset, ~50k samples

set	train	val	test
samples	31068	10187	8454

Classifier
Inception-v3
(very deep) CNN

Reg-1 Inception-v3 → (a_1, b_1)

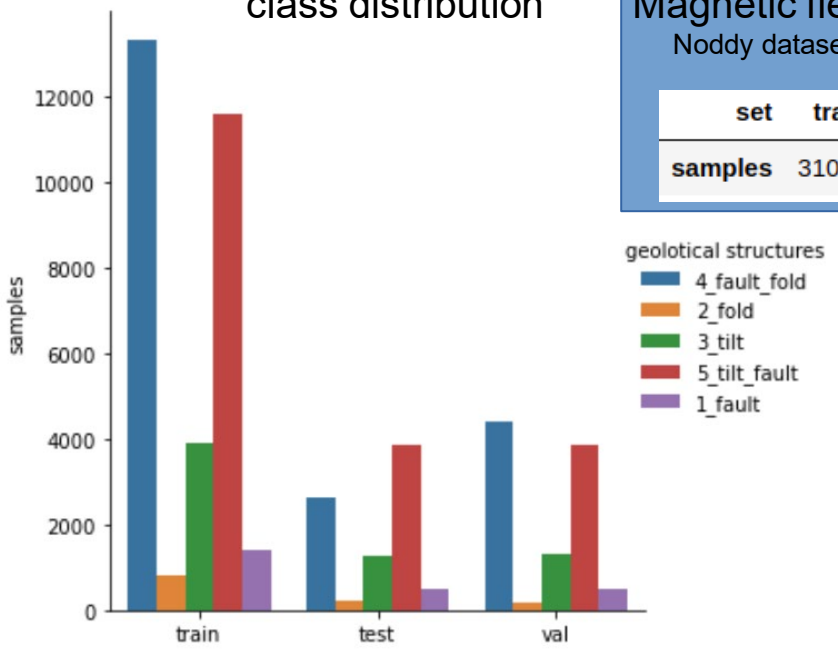
Reg-2 Inception-v3 → (a_1, b_1)

...

Reg-5 Inception-v3 → (a_1, b_1, a_2, b_2)

Guo's classifier

class distribution



Magnetic field 2D images

Noddy dataset, ~50k samples

set	train	val	test
samples	31068	10187	8454

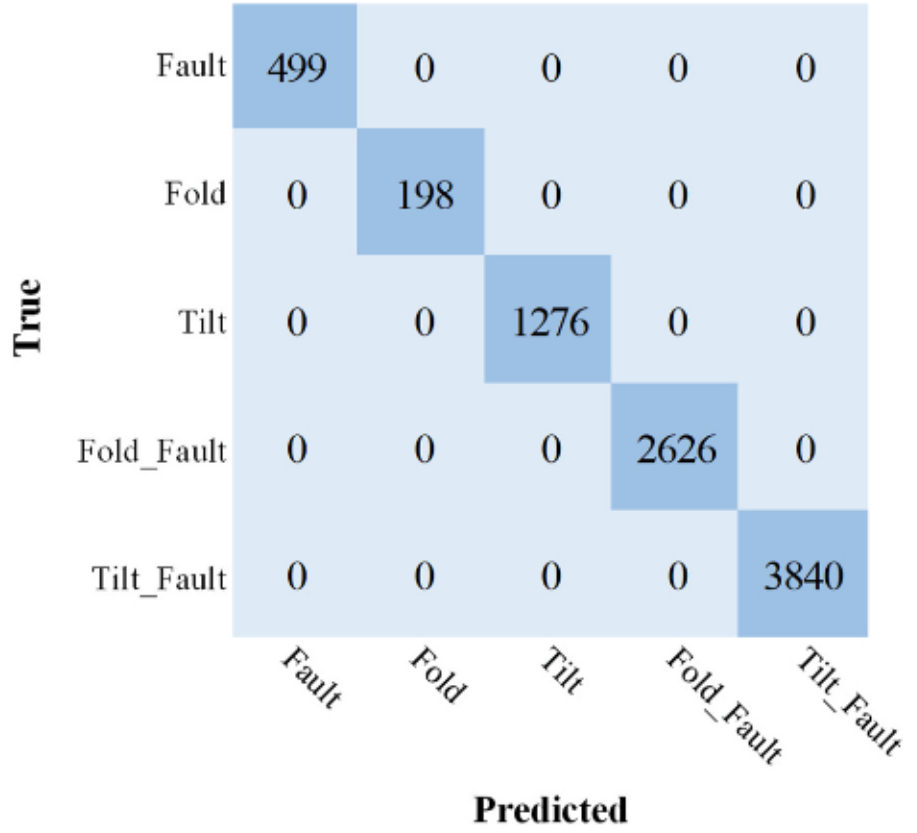
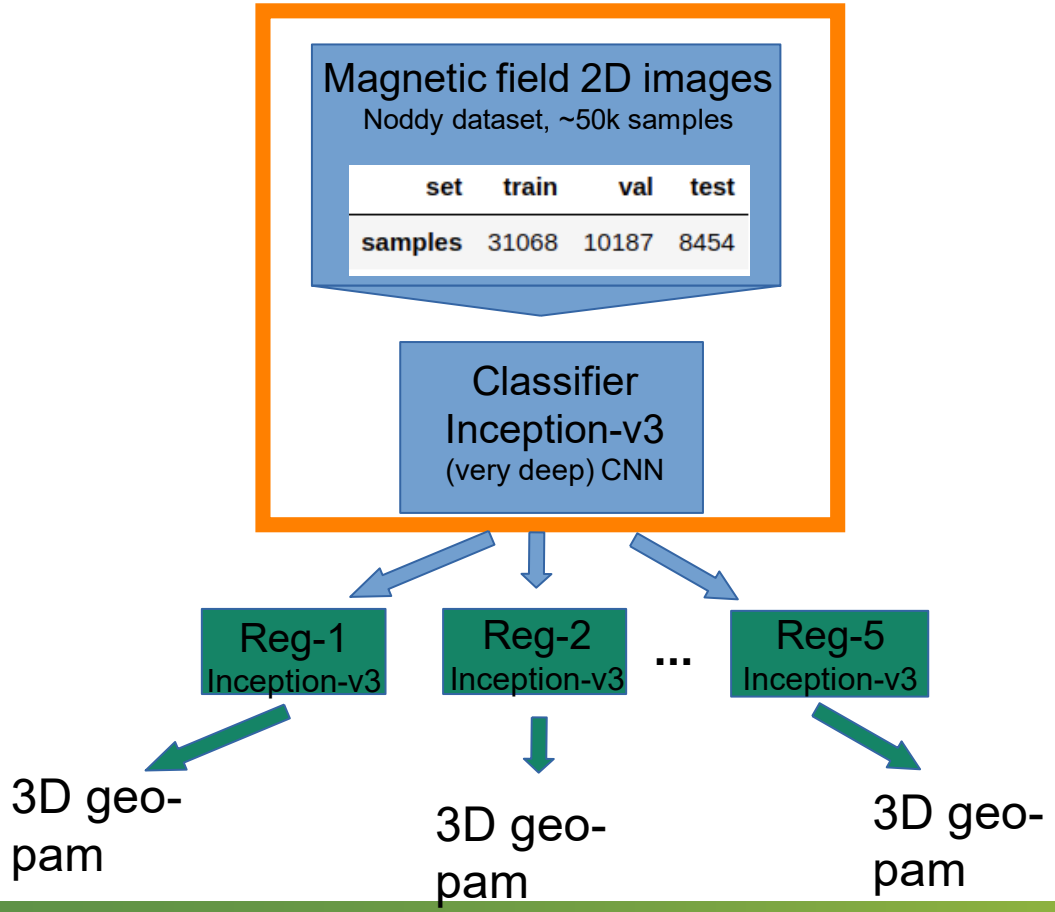


Fig. 4. Confusion matrix of the classification results when the model is applied to the test dataset.

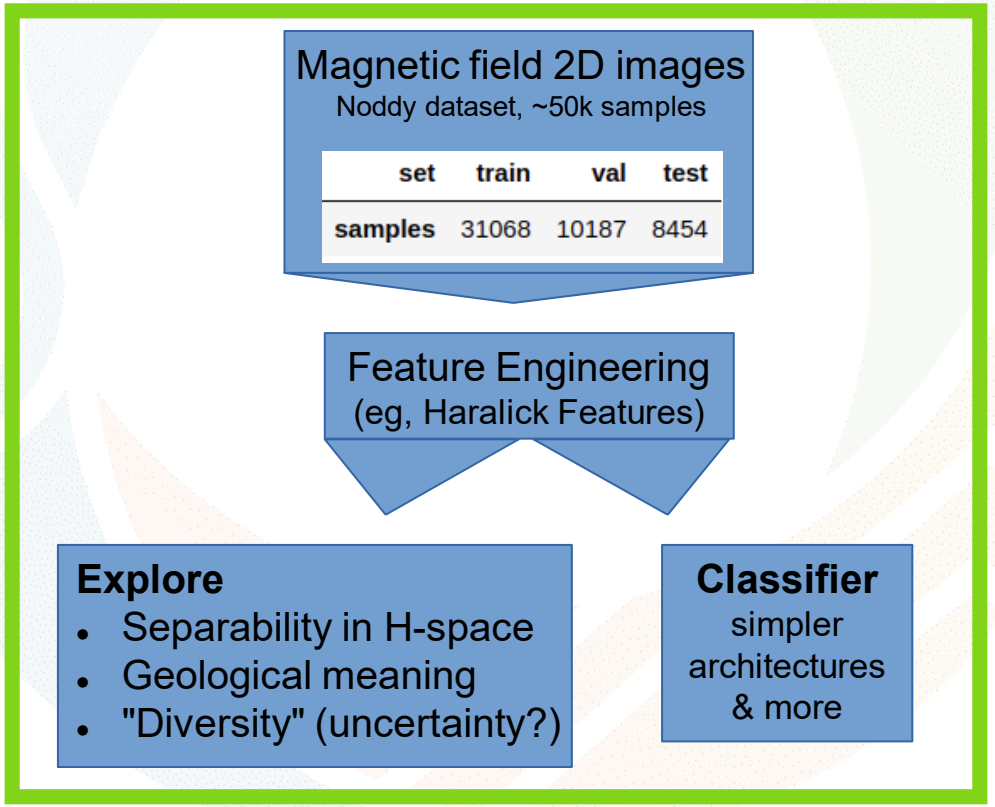
Guo et al, 2021

Current work

Guo et al 2021



Current work

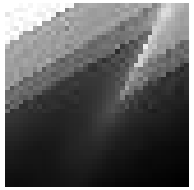


High dimensionality

Haralick Features

High dimensionality | encoding through texture features

Geophysics **mD vector**

tilt_fault 

$[f_1, f_2, \ldots, f_m]$

$NxN \rightarrow m, \quad m << N^2$

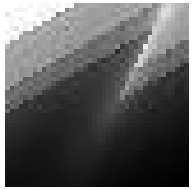
What can we explore and expect to achieve?

- Separability
- Geological meaning
- "Diversity" (class hierarchy?)
- Uncertainty

- Simpler ML architectures: SVM, KNN, RF
- Tasks: classification, regression (parameter extraction), feature ranking
- Scalabitliy (>>50k samples, less time/memory)
- Interpretability, explainability

High dimensionality | encoding through texture features

Geophysics | **mD vector**

tilt_fault  → $[f_1, f_2, \ldots, f_{13}]$

40x40 → 13, 13 << 1600

Grey Level Co-occurrence Matrix

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-3, NO. 6, NOVEMBER 1973

Textural Features for Image Classification

ROBERT M. HARALICK, K. SHANMUGAM, AND ITS'HAK DINSTEIN

III. Applications of Textural Features for Image Classification

In this section we present the results of our studies on the usefulness of the textural features for categorizing images. Three data sets were used in our study. These data sets were extracted from photomicrographs of different rocks, from aerial photographs of man-made and natural scenes, and from high-altitude satellite pictures of the earth. A brief description of the data sets and classification algorithms used and the results of classification experiments will be presented. For further details the interested reader is referred to [16]–[18].

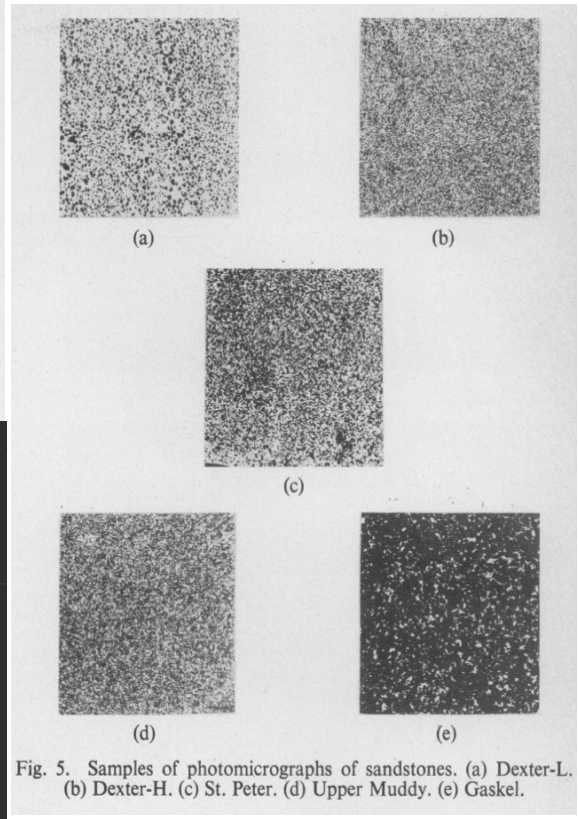


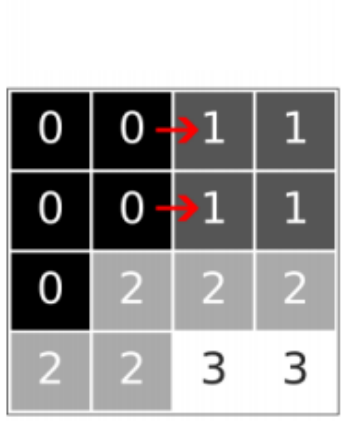
Fig. 5. Samples of photomicrographs of sandstones. (a) Dexter-L. (b) Dexter-H. (c) St. Peter. (d) Upper Muddy. (e) Gaskel.

Haralick *et al*, 1973

Grey Level Co-occurrence Matrix

Example: right (east) spatial relationship

Image



Framework matrix

neighbour pixel value -> / ref pixel value: ↓	0	1	2	3
0	0,0	0,1	0,2	0,3
1	1,0	1,1	1,2	1,3
2	2,0	2,1	2,2	2,3
3	3,0	3,1	3,2	3,3

GLCM

2	2	1	0
0	2	0	0
0	0	3	1
0	0	0	1

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}}$$

(normalized) GLCM

.166 (4/24)	.083 (2/24)	.042 (1/24)	0 (0/24)
.083	.166	0	0
.042	0	.250	.042
0	0	.042	.083

(normalized) GLCM ~ Joint probability distribution of two pixels having the same grey level

Hall-Beyer, 2007. GLCM Texture: A Tutorial v. 1.0 through 2. http://hdl.handle.net/1880/51900

Texture measures (features)

- Summarize the GLCM in helpful ways
- Several of them (13-17?): allows choice to fit the problem at hand
- Most of them are weighted averages of the GLCM

1. A. Contrast (this is also called "sum of squares variance" and occasionally “inertia”):

$$\sum_{i,j=0}^{N-1} P_{i,j}\left(i-j\right)^2$$

Contrast equation

Explanation:

When i and j are equal, the cell is on the diagonal and (i-j)=0. These values represent pixels entirely similar to their neighbour, so they are given a weight of 0 (no contrast).
If i and j differ by 1, there is a small contrast, and the weight is 1.
If i and j differ by 2, contrast is increasing and the weight is 4.
The weights continue to increase exponentially as (i-j) increases.

Hall-Beyer, 2007. GLCM Texture: A Tutorial v. 1.0 through 2. http://hdl.handle.net/1880/51900

SCIENTIFIC REPORTS

Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters

Patrik Brynolfsson[1], David Nilsson[2], Turid Torheim[3], Thomas Asklund[1], Camilla Thellenberg Karlsson[1], Johan Trygg[2], Tufve Nyholm[1] & Anders Garpebring[1]

In recent years, texture analysis of medical images has become increasingly popular in studies investigating diagnosis, classification and treatment response assessment of cancerous disease. Despite numerous applications in oncology and medical imaging in general, there is no consensus regarding texture analysis workflow, or reporting of parameter settings crucial for replication of results. The aim of this study was to assess how sensitive Haralick texture features of apparent diffusion coefficient (ADC) MR images are to changes in five parameters related to image acquisition and pre-processing: noise, resolution, how the ADC map is constructed, the choice of quantization method, and the number of gray levels in the quantized image. We found that noise, resolution, choice of quantization method and the number of gray levels in the quantized images had a significant influence on most texture features, and that the effect size varied between different features. Different methods for constructing the ADC maps did not have an impact on any texture feature. Based on our results, we recommend using images with similar resolutions and noise levels, using one quantization method, and the same number of gray levels in all quantized images, to make meaningful comparisons of texture feature results between different subjects.

Feature	Equation
Autocorrelation	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i \cdot j)p(i,j)$
Cluster Prominence	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i+j-2\mu)^3 p(i,j)$
Cluster shade	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i+j-2\mu)^4 p(i,j)$
Contrast	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i-j)^2 p(i,j)$
Correlation	$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(i \cdot j)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$
Difference entropy	$-\sum_{k=0}^{N-1}p_{x-y}(k)\log p_{x-y}(k)$
Difference variance	$\sum_{k=0}^{N-1}(k-\mu_{x-y})^2 p_{x-y}(k)$
Dissimilarity	$\sum_{i=1}^{N}\sum_{j=1}^{N}\lvert i-j\rvert \cdot p(i,j)$
Energy	$\sum_{i=1}^{N}\sum_{j=1}^{N}p(i,j)^2$
Entropy	$-\sum_{i=1}^{N}\sum_{j=1}^{N}p(i,j)\log p(i,j)$
Homogeneity	$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p(i,j)}{1+(i-j)^2}$
Information measure of correlation 1	$\frac{HXY-HXY1}{\max(HX,HY)}$
Information measure of correlation 2	$\sqrt{1-\exp[-2(HXY2-HXY)]}$
Inverse difference	$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p(i,j)}{1+\lvert i-j\rvert}$
Maximum probability	$\max_{i,j}p(i,j)$
Sum average, μ_{x+y}	$\sum_{k=2}^{2N}kp_{x+y}(k)$
Sum entropy	$-\sum_{k=2}^{2N}p_{x+y}(k)\log p_{x+y}(k)$
Sum of squares	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i-\mu)^2 p(i,j)$
Sum variance	$\sum_{k=2}^{2N}(k-\mu_{x+y})^2 p_{x+y}(k)$

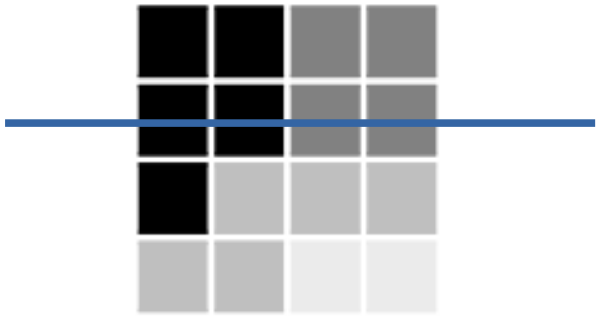
Table 3. Haralick texture features calculated from GLCMs. There was an error in the definition of *Sum variance* in Haralick *et al.*[1], which has been corrected.

Brynolfsson et al, 2017

Rotational "*invariance*"

Right → Horizontal

$A_{horizontal} = A + A^T$



$\vec{f}_{horiz} = [f_1, f_2, \ldots, f_{13}]$

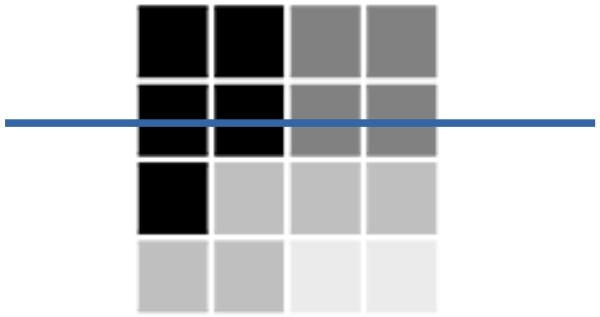
The various features which we suggest are all functions of distance and angle. The angular dependencies present a special problem. Suppose image *A* has features *a*, *b*, *c*, and *d* for angles 0°, 45°, 90°, and 135°, respectively, and image *B* is identical to *A* except that *B* is rotated 90° with respect to *A*. Then *B* will have features *c*, *d*, *a*, and *b* for angles 0°, 45°, 90°, and 135°, respectively. Since the texture context of *A* is the same as the texture context of *B*, any decision rule using the angular features *a*,*b*,*c*,*d* must produce the same results for *c*,*d*,*a*,*b*. To guarantee this, we suggest that the angularly dependent features not be used directly. Instead we suggest that two functions of *a*, *b*, *c*, and *d*, their average and range (which are invariant under rotation), be used as inputs to the classifier.

Haralick *et al*, 1973

Images → feature vectors

Right → Horizontal

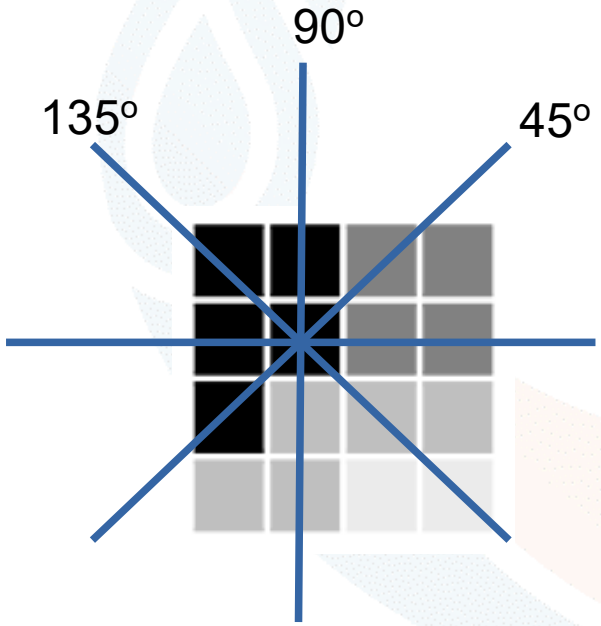
$$A_{horizontal} = A + A^T$$



$$\vec{f}_{horiz} = [f_1, f_2, \ldots, f_{13}]$$

4 Directions

- 4 GLCM in total
- 4x13 features
- Average of all directions → approximate rotational invariant feature



$$\vec{f}_0 = [f_1, f_2, \ldots, f_{13}]$$

$$\vec{f}_{45} = [f_1, f_2, \ldots, f_{13}]$$

$$\vec{f}_{90} = [f_1, f_2, \ldots, f_{13}]$$

$$\vec{f}_{135} = [f_1, f_2, \ldots, f_{13}]$$

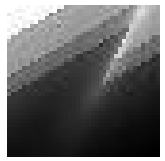
$$\vec{f}_{av} = [\langle f_1 \rangle_{av}, \ldots, \langle f_{13} \rangle_{av}]$$

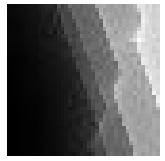
High dimensionality | encoding through texture features

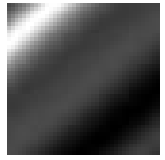
40x40 → 13, 13<< 1600

Geophysics

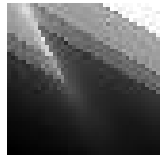
ASM Contrast HXY2

tilt_fault  [0.00101, 332.10, …, 0.999884]

tilt_fault  [0.00212, 212.44, …, 0.999868]

fault_fold  [0.00298, 84.02, …, 0.999728]

...

tilt_fault  [0.00101, 332.08, …, 0.999885]

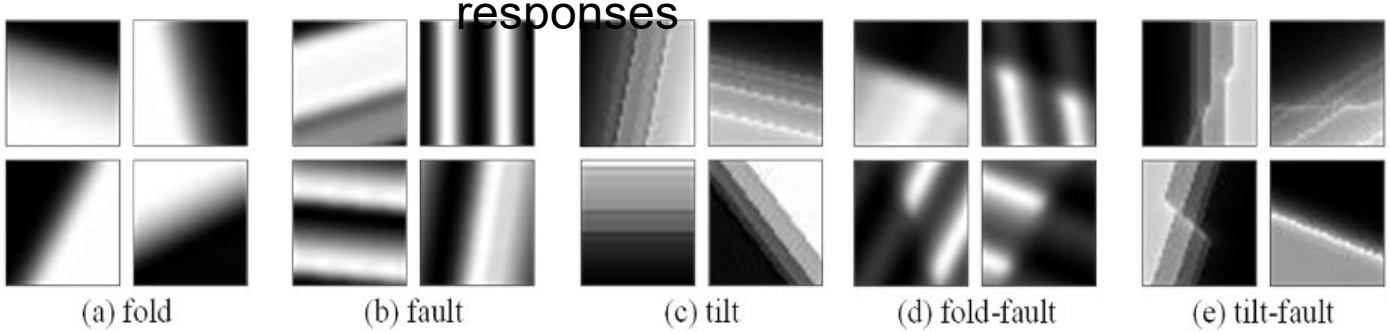
List of features used

1. ASM, Angular second moment
2. Contrast
3. Correlation
4. SSV, Sum of squares: variance
5. IDM, Inverse difference moment
6. SA, sum average
7. SV, sum variance
8. Sum Entropy
9. Entropy
10. DF, Difference variance
11. DE, Difference Entropy
12. and 13, Information Measures correlations:
    * HXY1
    * HXY2

Higher similarity

And now… What?

2D mag responses



New representation

f_1	f_2	f_3	…	f_13	geo_str
0.014681	47.366432	0.996898		0.999906	1_fault
0.022420	32.698278	0.992737		0.999925	1_fault
0.022604	91.225263	0.995584		0.999948	1_fault
0.062451	153.235096	0.993851		0.999567	1_fault
0.067190	99.367636	0.995896	…	0.999750	1_fault
…	…	…		…	…
0.039720	248.273895	0.975236		0.999845	5_tilt_fault
0.005483	266.549269	0.982171		0.999948	5_tilt_fault
0.003695	283.829787	0.979010		0.999809	5_tilt_fault
0.026471	410.904516	0.975690		0.999693	5_tilt_fault
0.010526	337.019157	0.979102	…	0.999716	5_tilt_fault

31050 rows × 14 columns

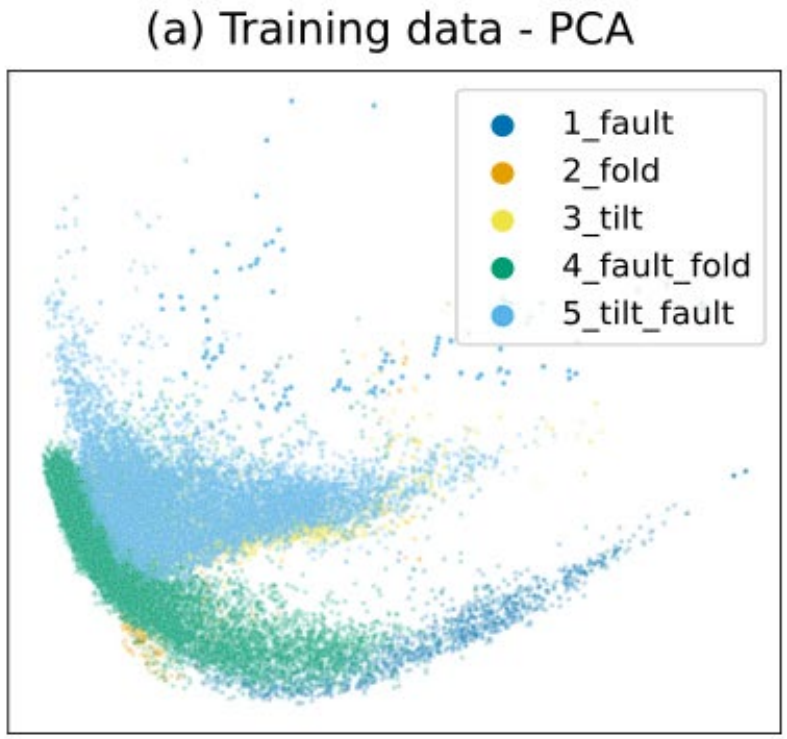
We now have a 13D space to represent the geophysics. Is it meaningful?

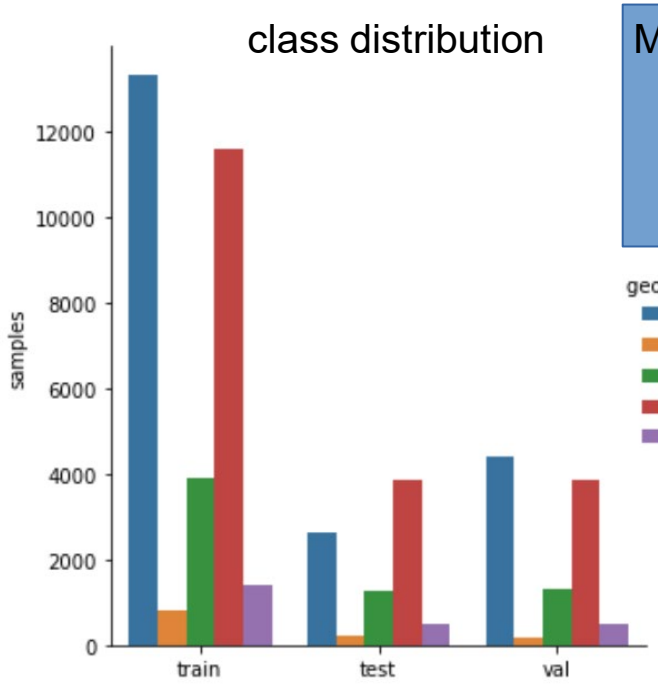
- Data **separability**?
  - **Simpler** ML architectures?
- 13D space has a **geological interpretation**?
- Projection of **real-world data**?

Exploring separability

(in the “Haralick space”)

Exploring separability



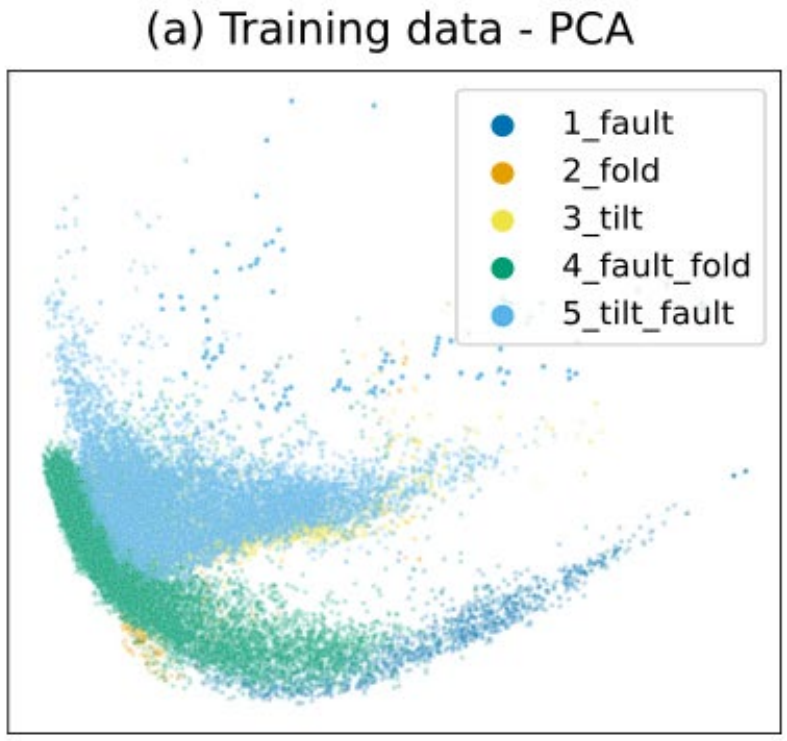


Magnetic field 2D images

Noddy dataset, ~50k samples

set	train	val	test
samples	31068	10187	8454

Exploring separability

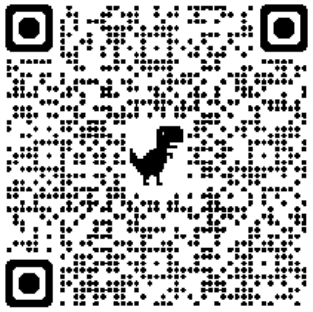


t-SNE projection

(t-distributed stochastic neighbor embedding)

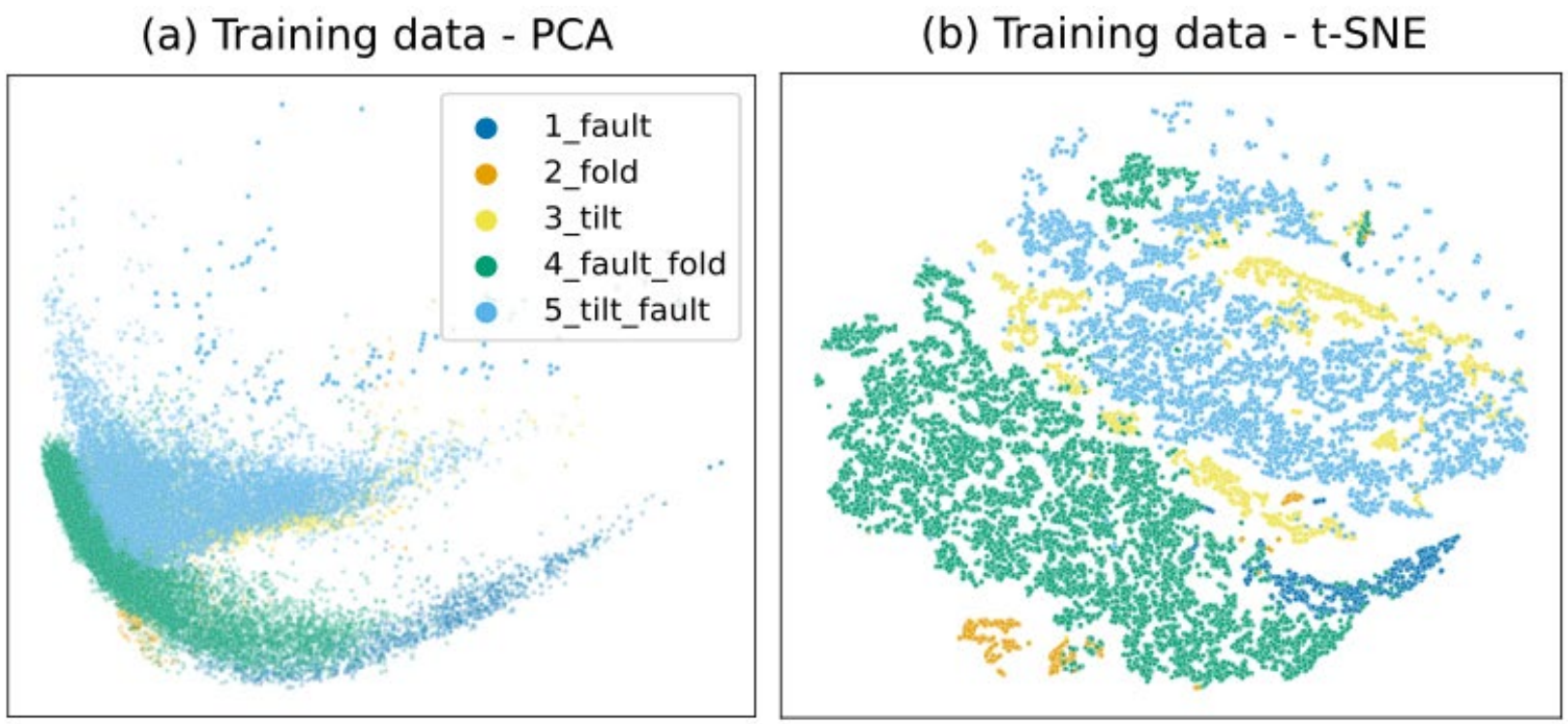
Other possible options*: Uniform Manifold Approximation and Projection (**UMAP**), Multidimensional scaling (**MDS**), Self-organizing map (**SOM**)

* and 22 others at Wikipedia

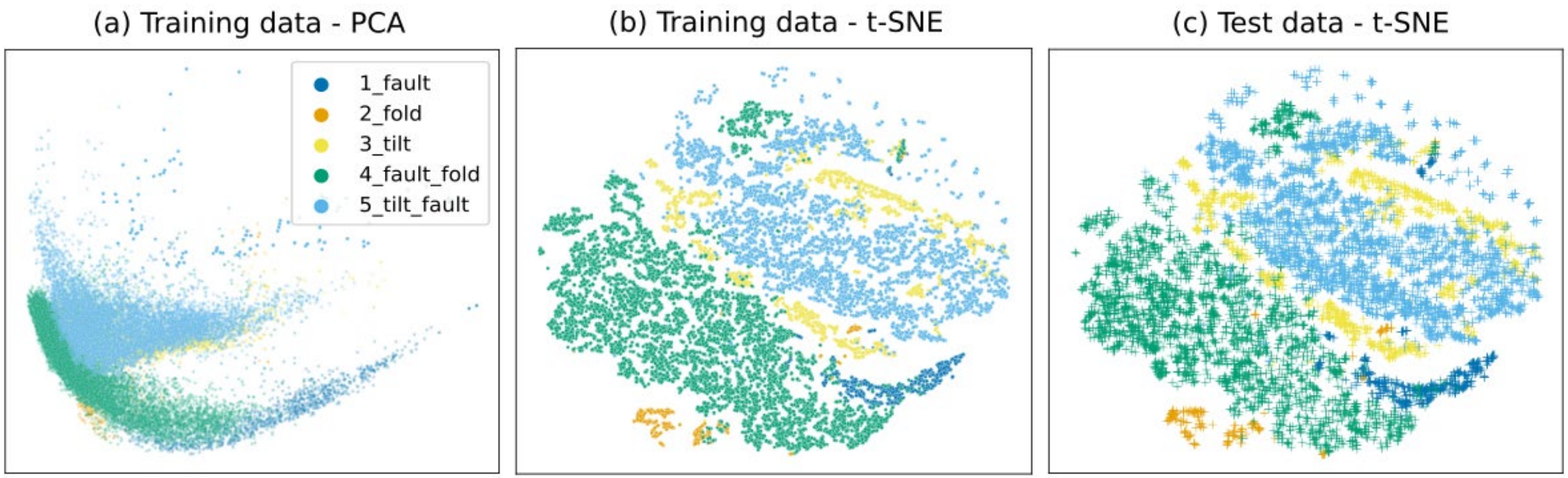


* Van der Maaten, Laurens, and Geoffrey Hinton. *Visualizing data using t-SNE*. Journal of machine learning research 9, no. 11 (2008).
* Kobak, D. & Berens, P. *The art of using t-SNE for single-cell transcriptomics*. Nat Commun 10, 5416 (2019).

Exploring separability



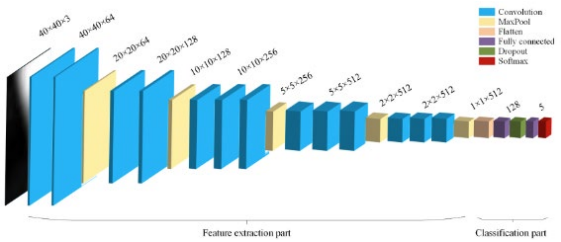
Exploring separability

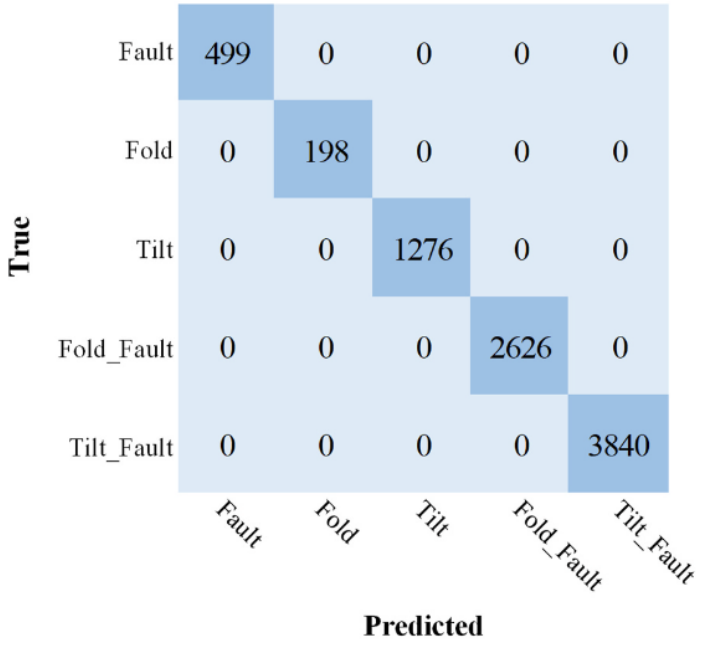


Simpler ML architectures

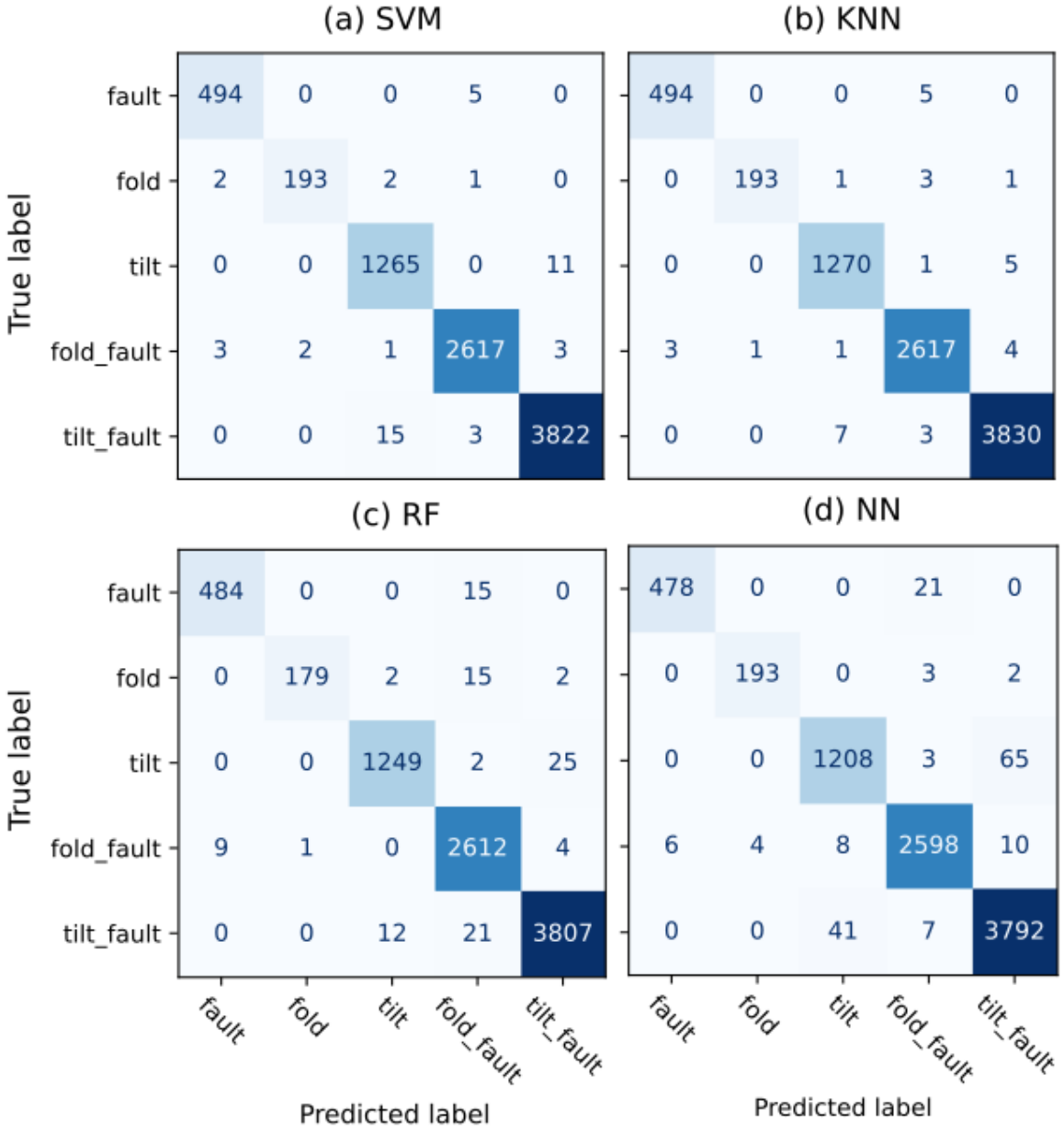
SVM, KNN, Random Forest

Classification | “Simple” ML architectures





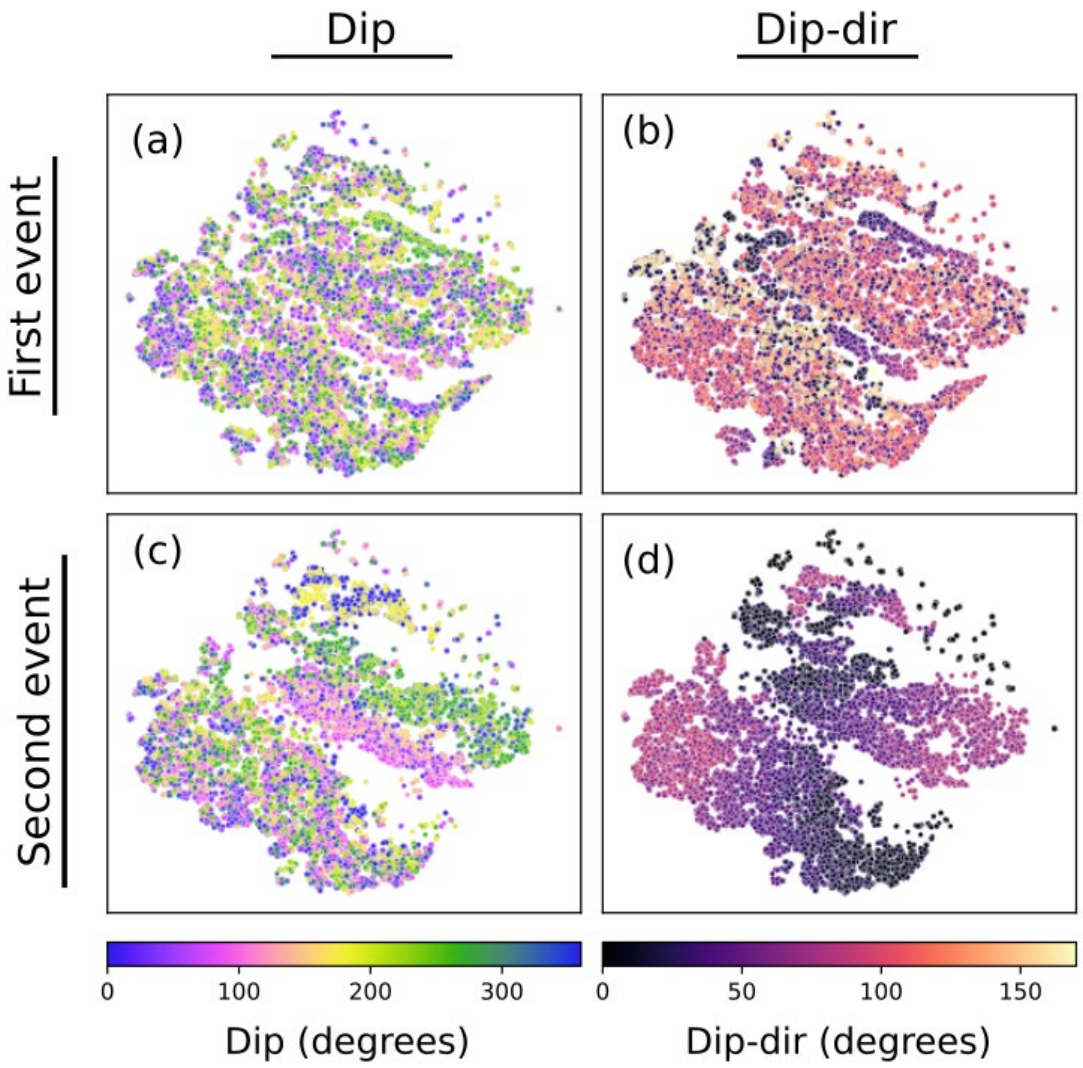
Guo *et al, 2021*



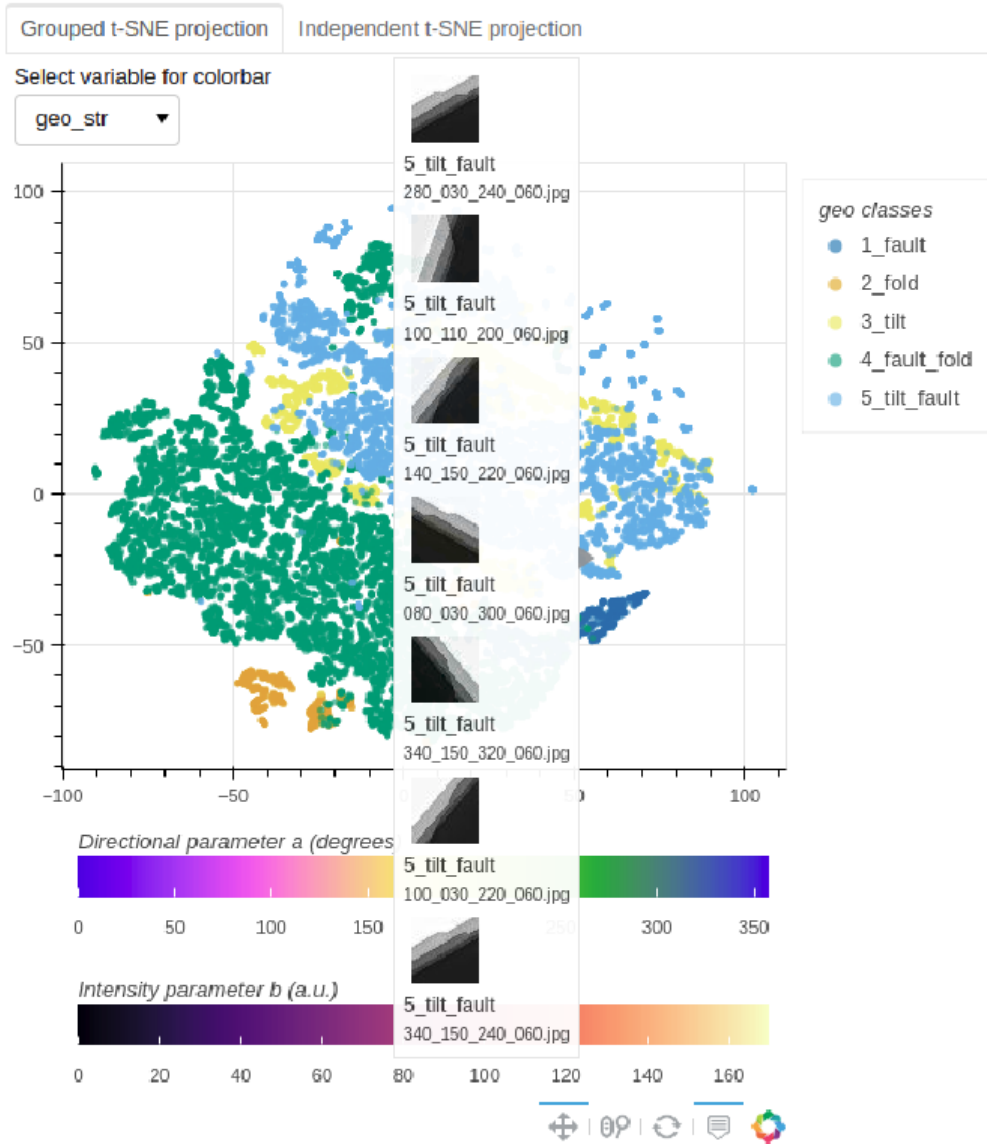
Geological Interpretation

(More @ 3-4 pm, NoddyVerse demo!)

Geological Interpretation

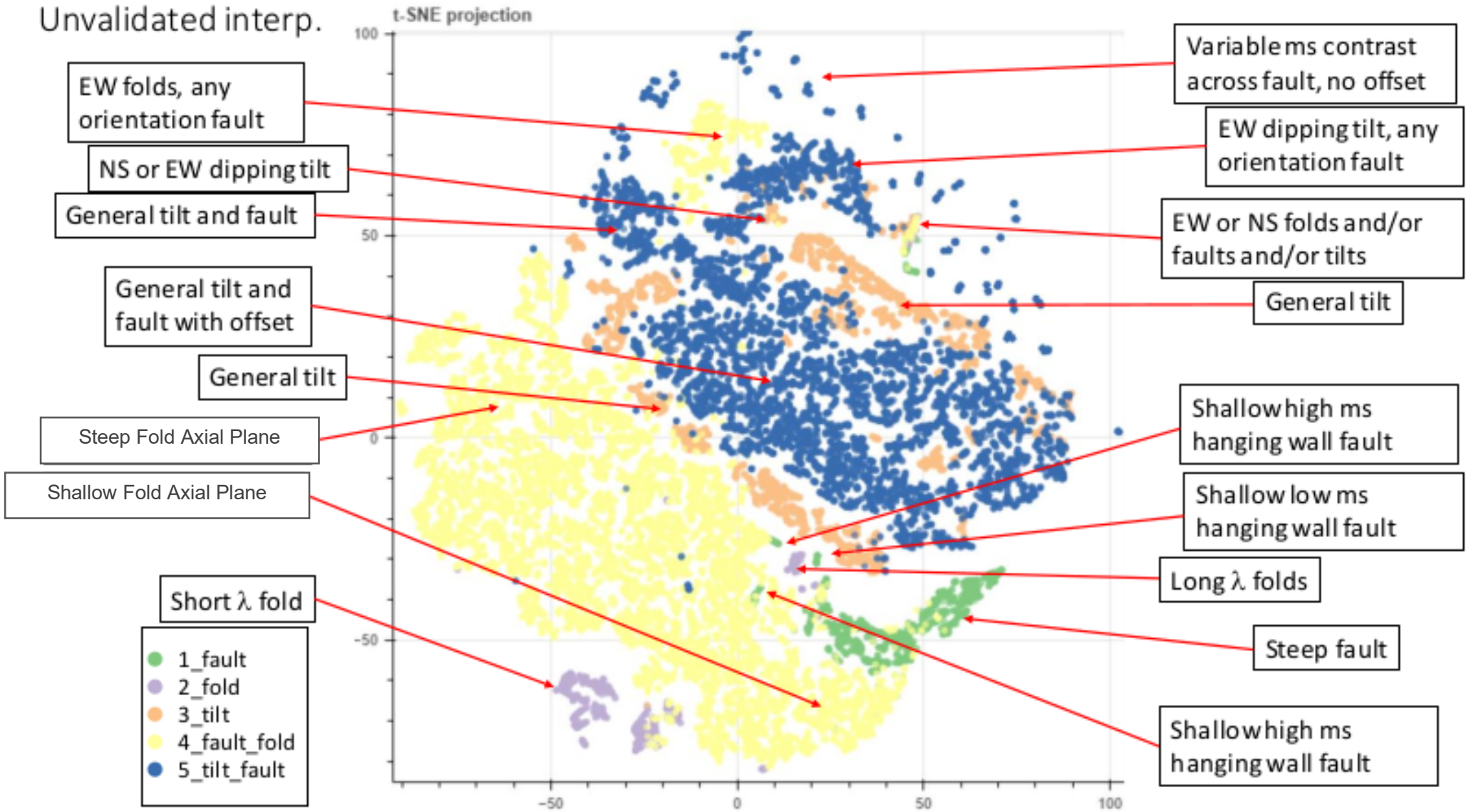


(More @ 3-4 pm, NoddyVerse demo!)



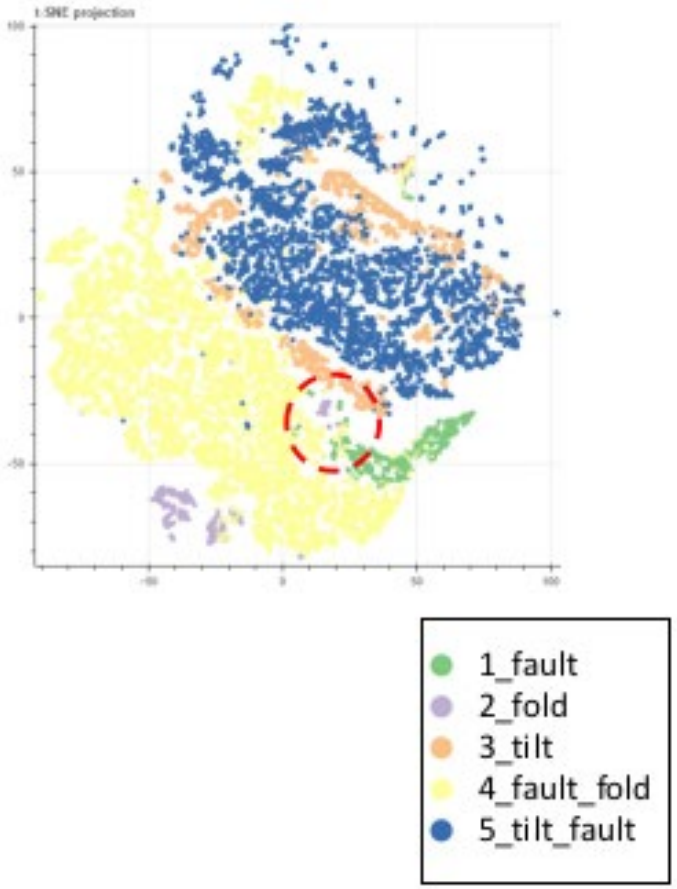
http://13.211.217.129/6IAS/integrated-data-vis%202.html

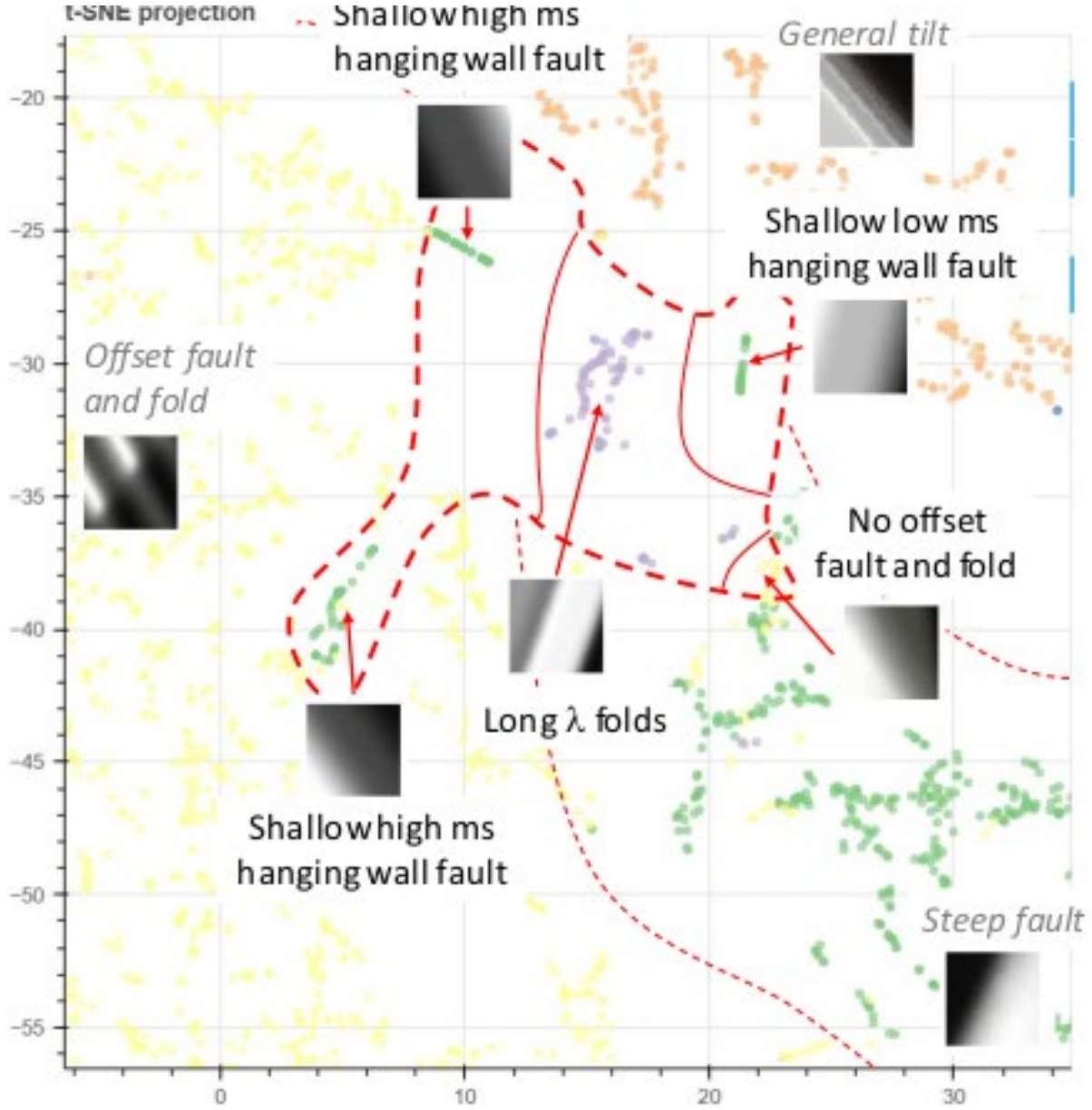
Geological Interpretation



Geological Interpretation

Mixed Zone: micro zones, different signatures

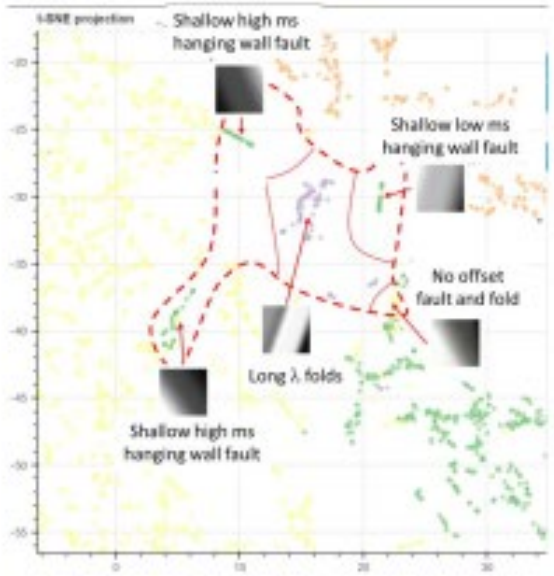


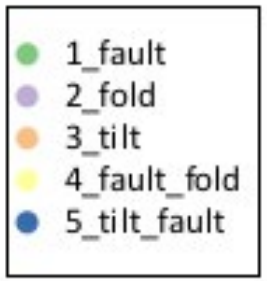


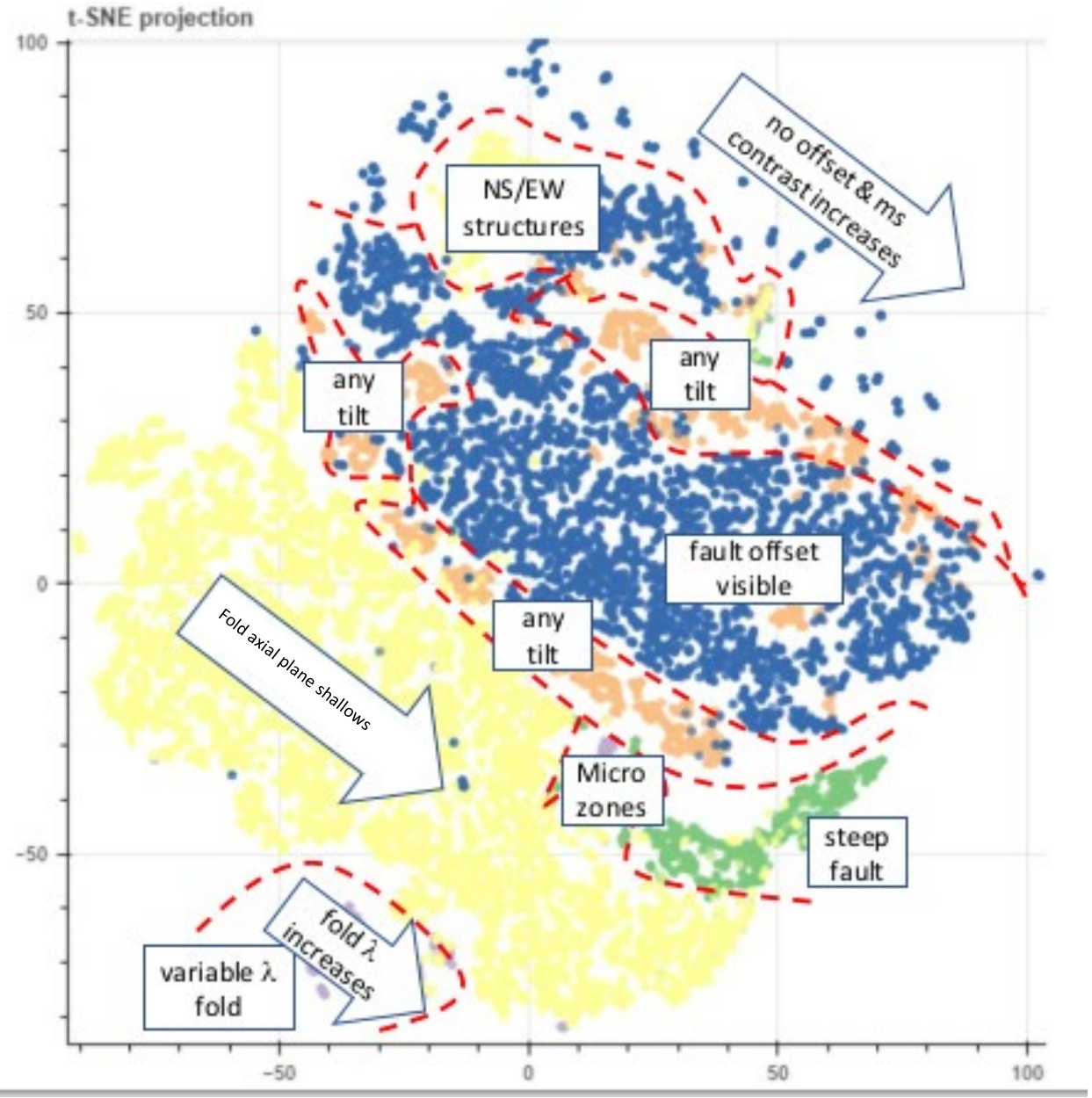
Geological Interpretation

Geological interpretation of t-SNE

Micro zones







Current and future work

Real-world data

Projection

NoddyVerse dataset (hands-on)

343 classes and 310K data points

“Maths”+Theory

Optimal (hyper) parameters, features and the curse of dimensionality

Current and future work

Real-world data

Projection

NoddyVerse dataset

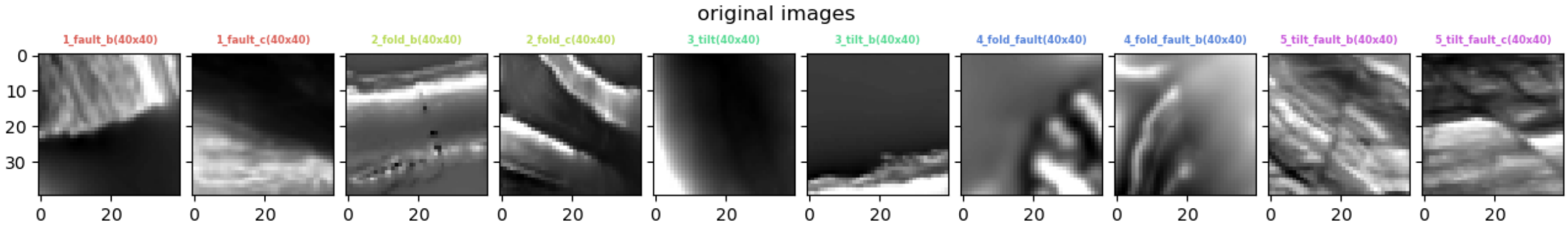
343 classes and 310K data points

“Maths”+Theory

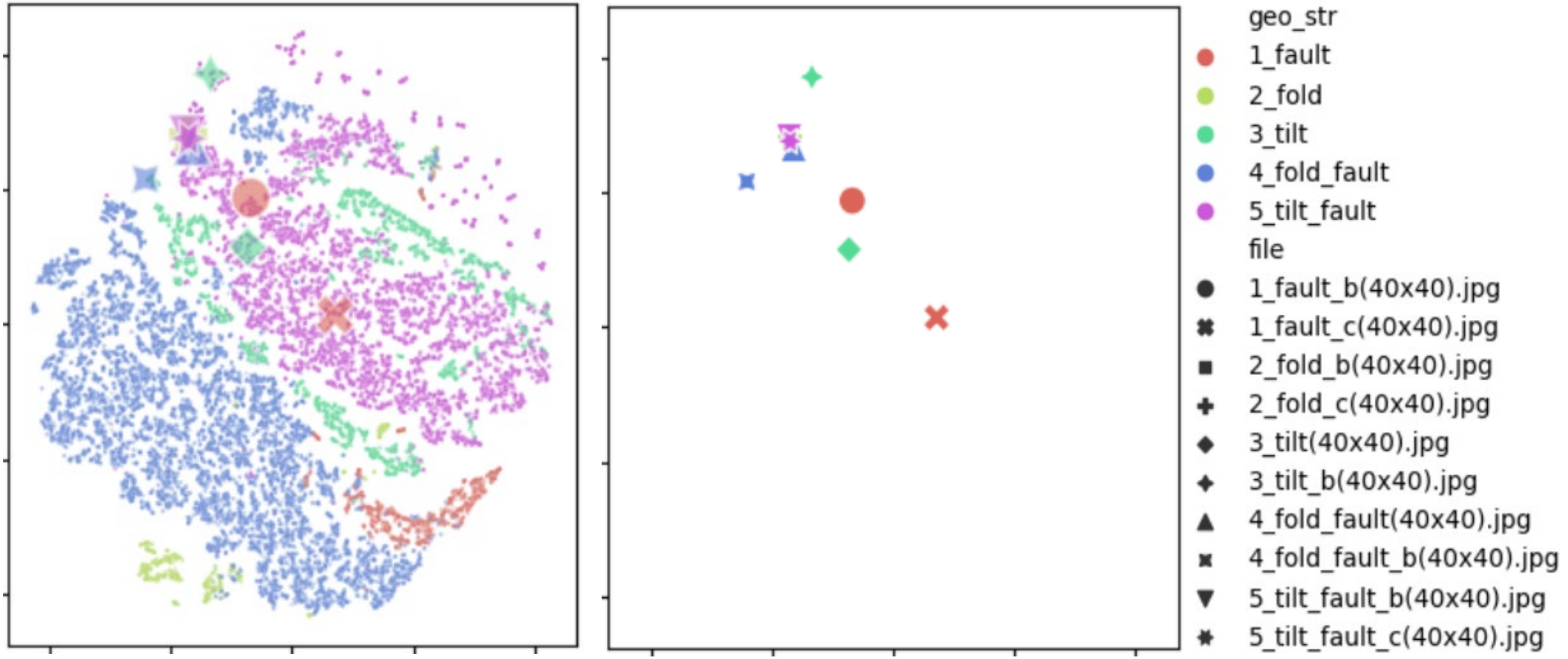
Optimal (hyper) parameters, features and the curse of dimensionality

Rear-world data provided by an expert

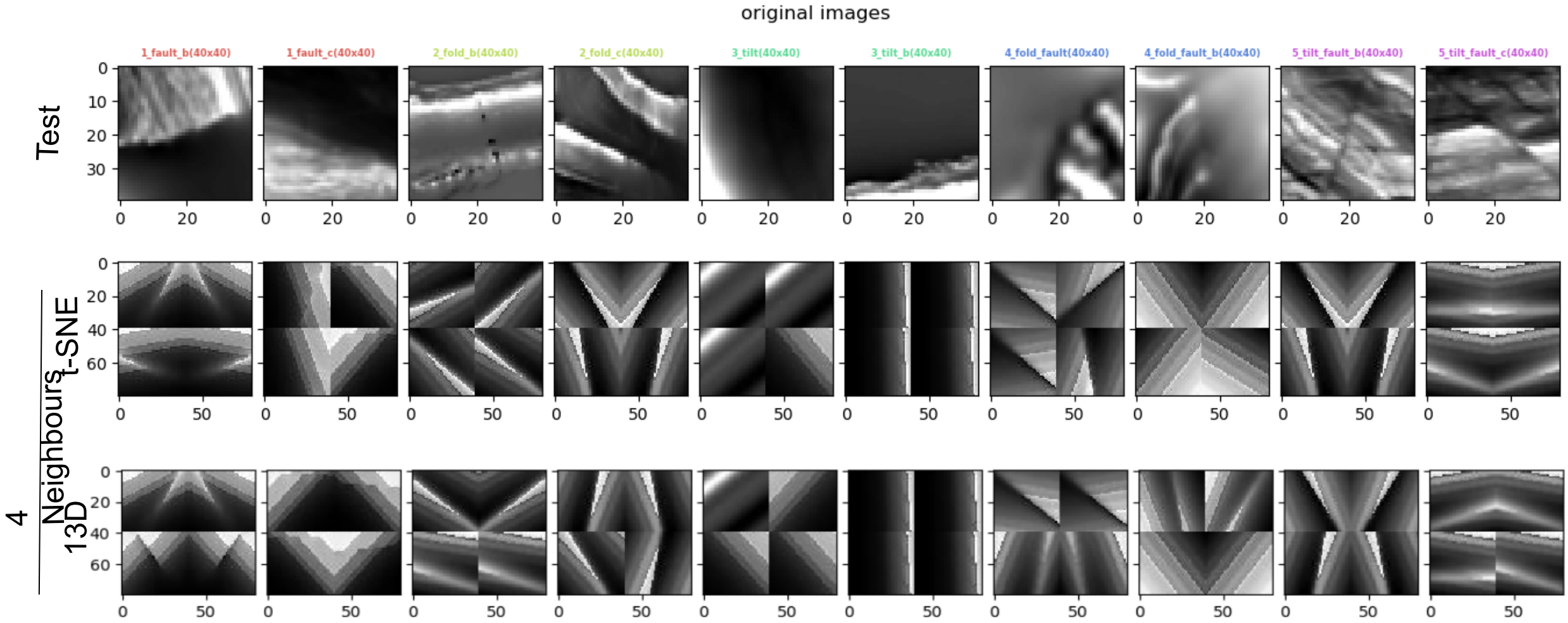
- Manually classified
- Two instances for each of the five classes investigated by Guo 2021, with similar geological scale
- Pre-processing: 600x600 → 40x40; Haralick features with distance = 1



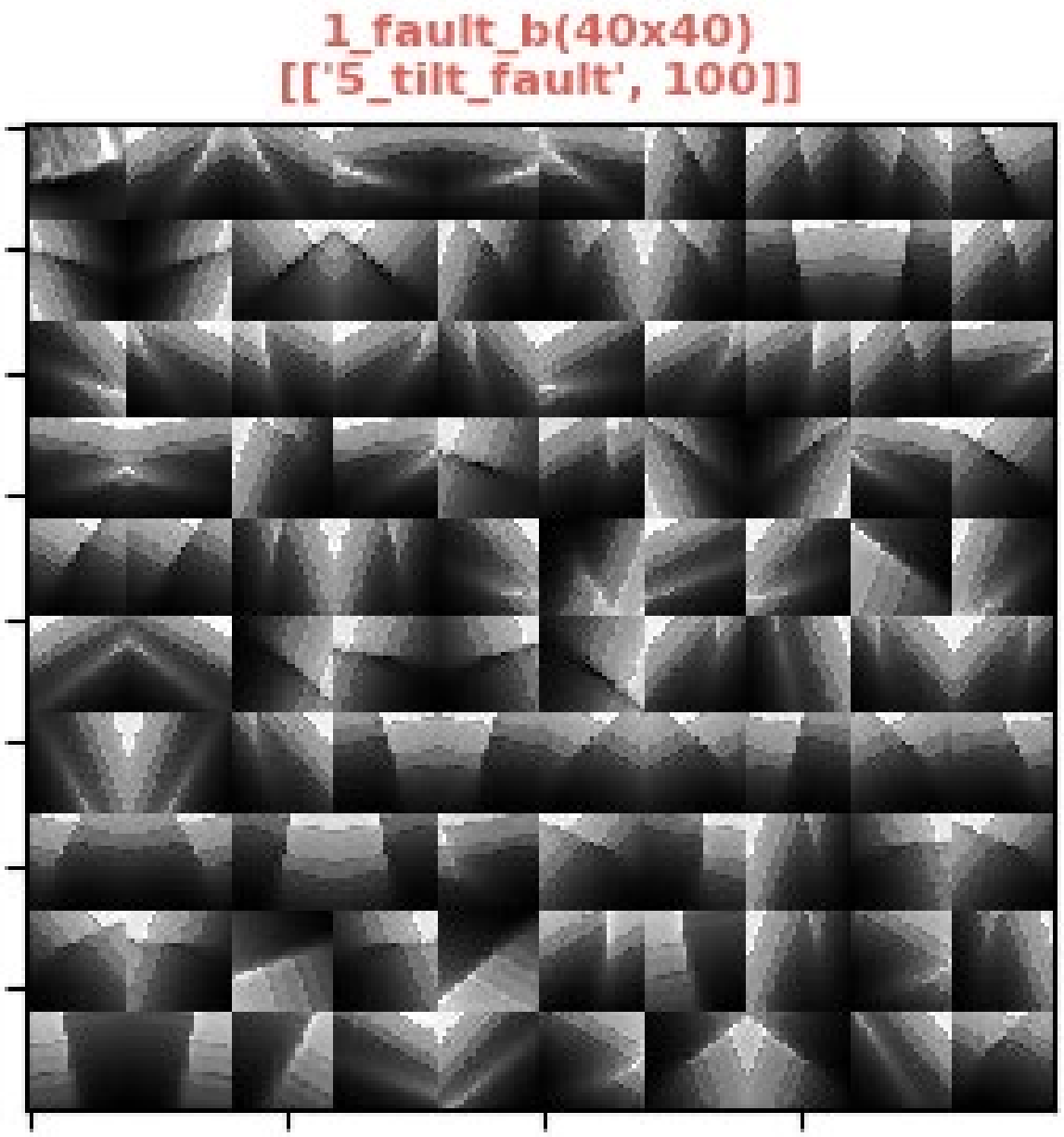
T-SNE | training data & projected new instances

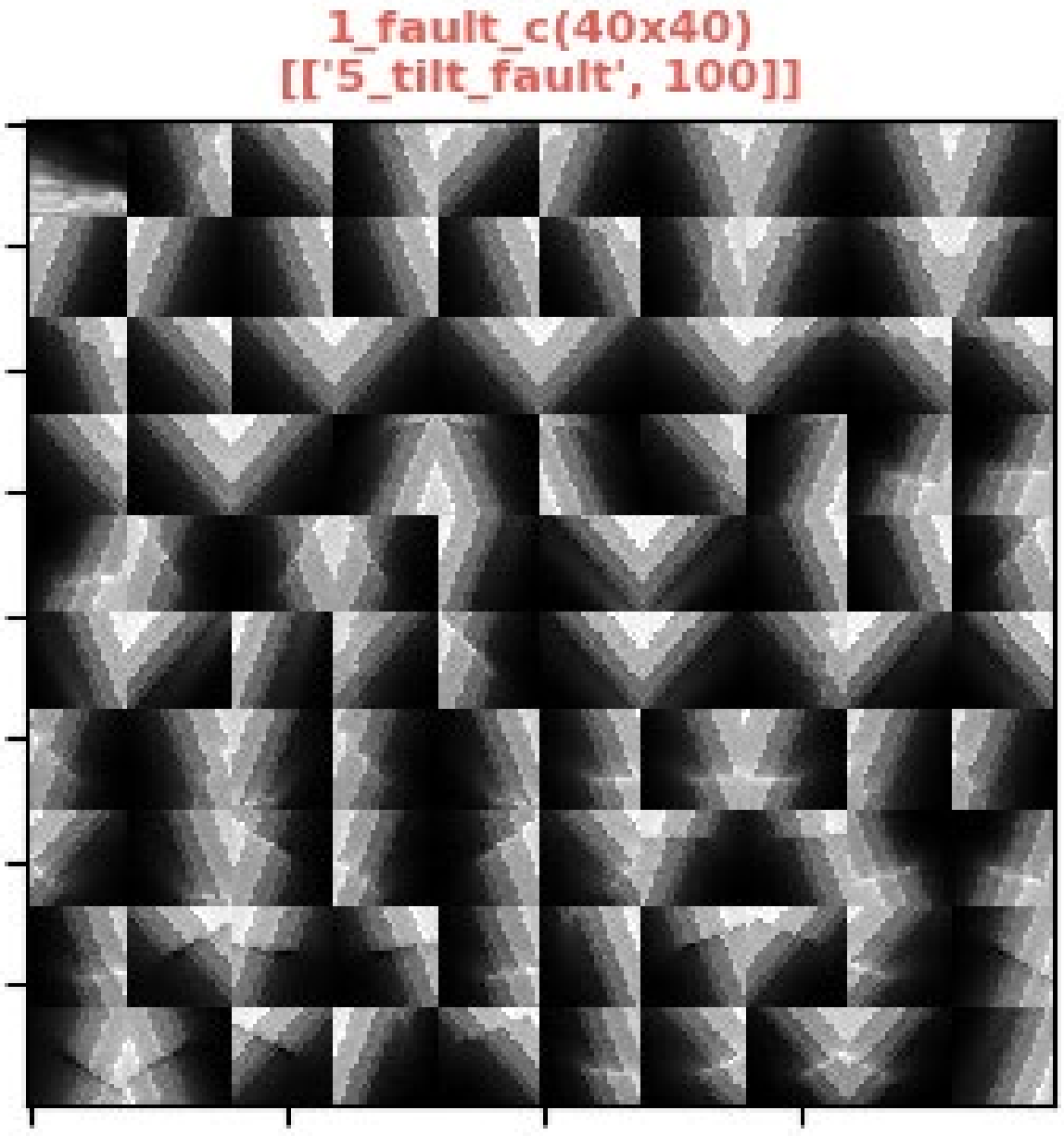


Four nearest neighbors (t-sne 2d-space and original 13d space)



1st tile is the test image (t-SNE neighbours)





Current and future work

Real-world data

Projection

NoddyVerse dataset (hands-on)

343 classes and 310K data points

“Maths”+Theory

Optimal (hyper) parameters, features and the curse of dimensionality

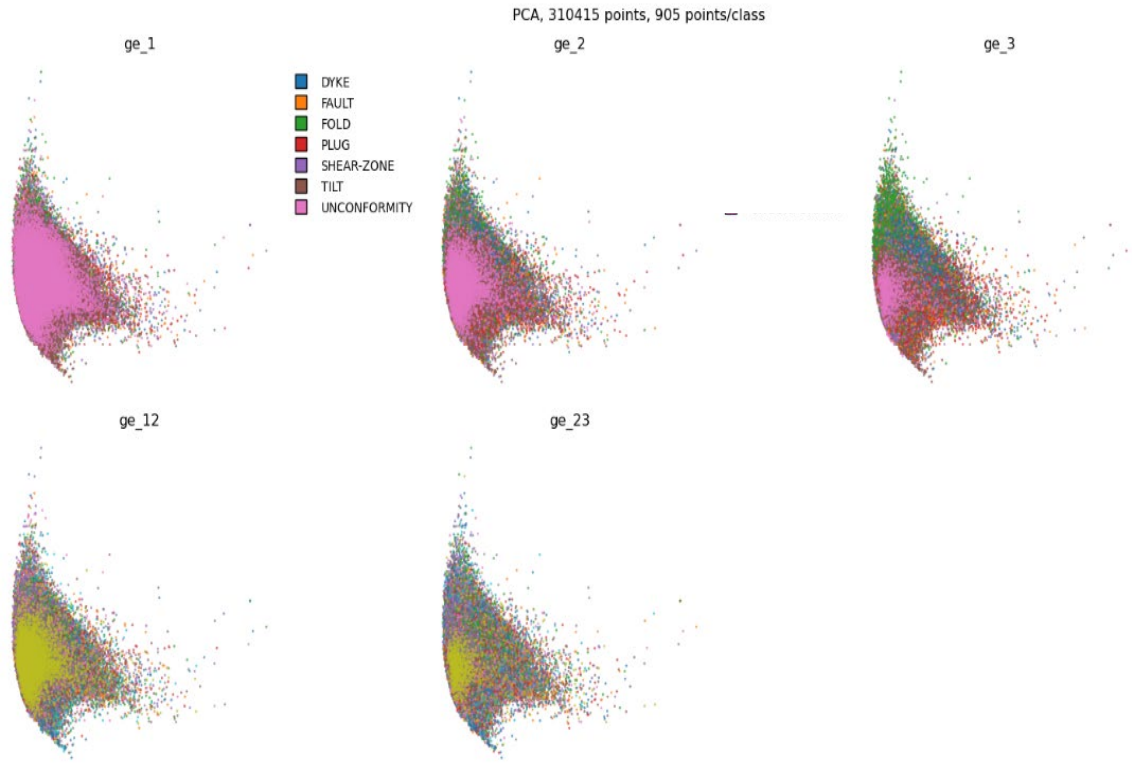
343 classes and 310K data points…

Encoded into only 13 dimensions...

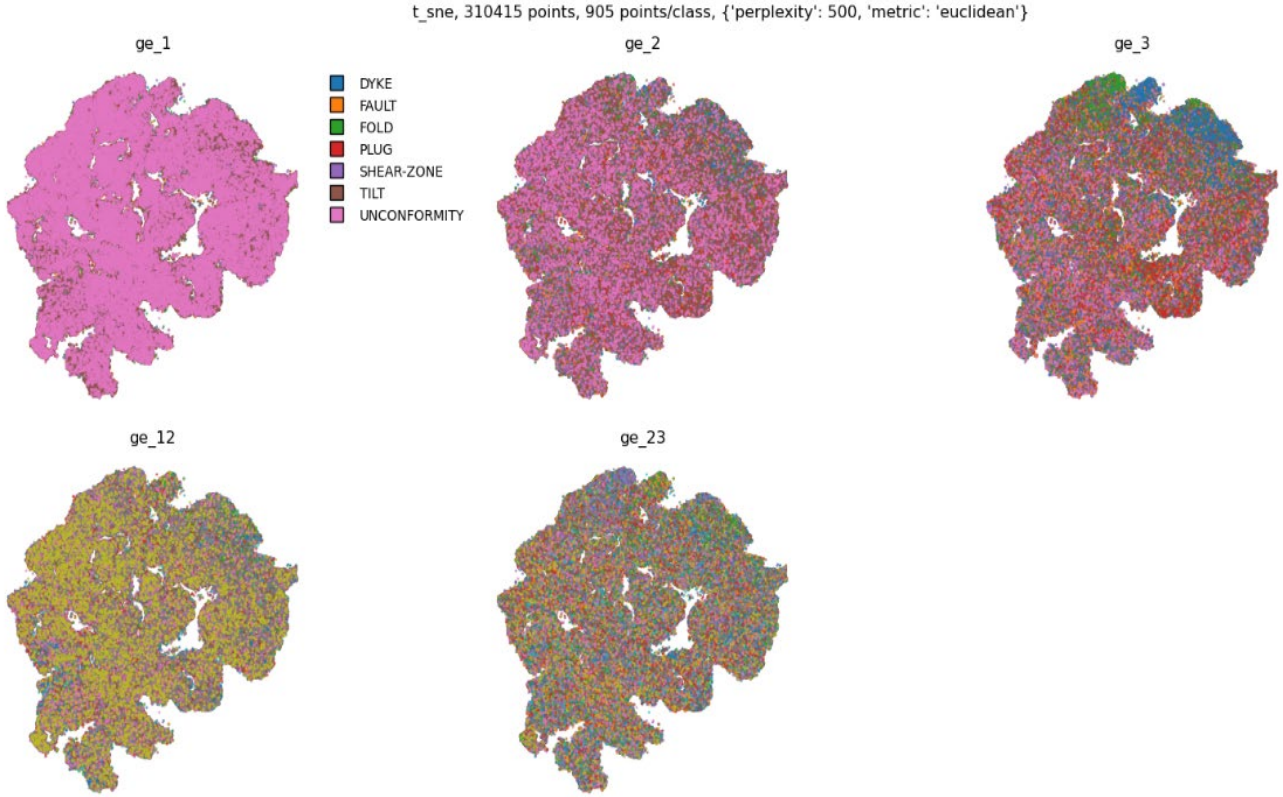
Visualized in only 2 dimensions...

Geological event history *[E1, E2, E3]* → color by E_i class

PCA



t-SNE



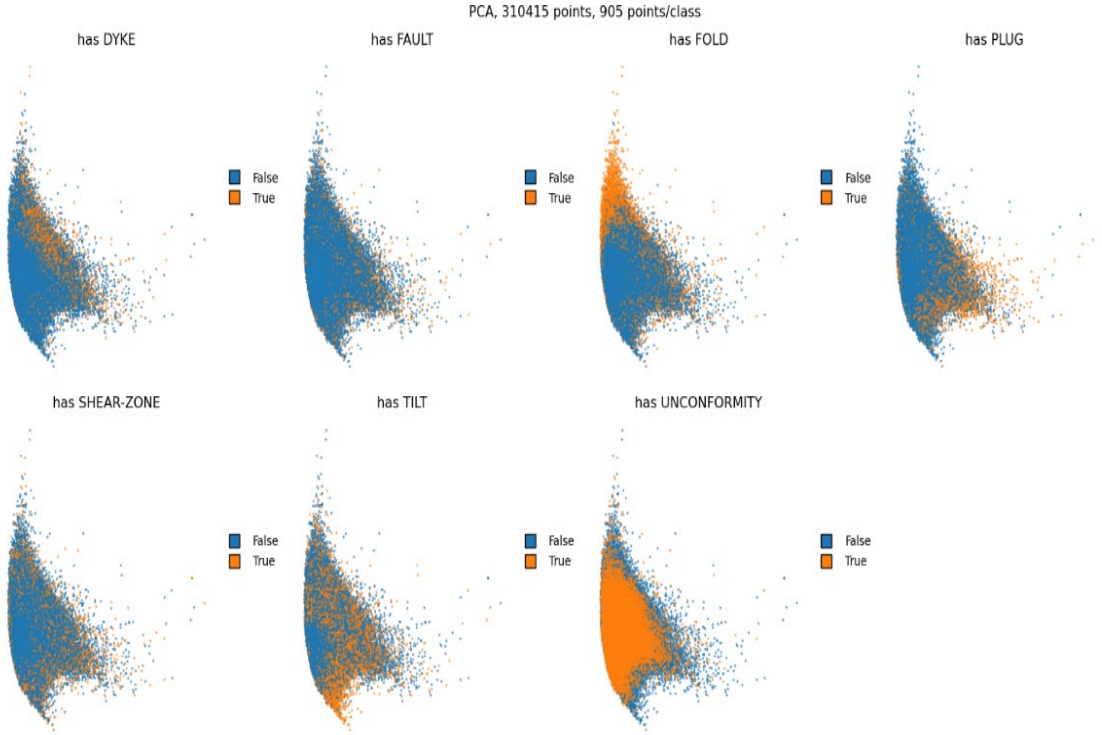
343 classes and 310K data points…

Encoded into only 13 dimensions...

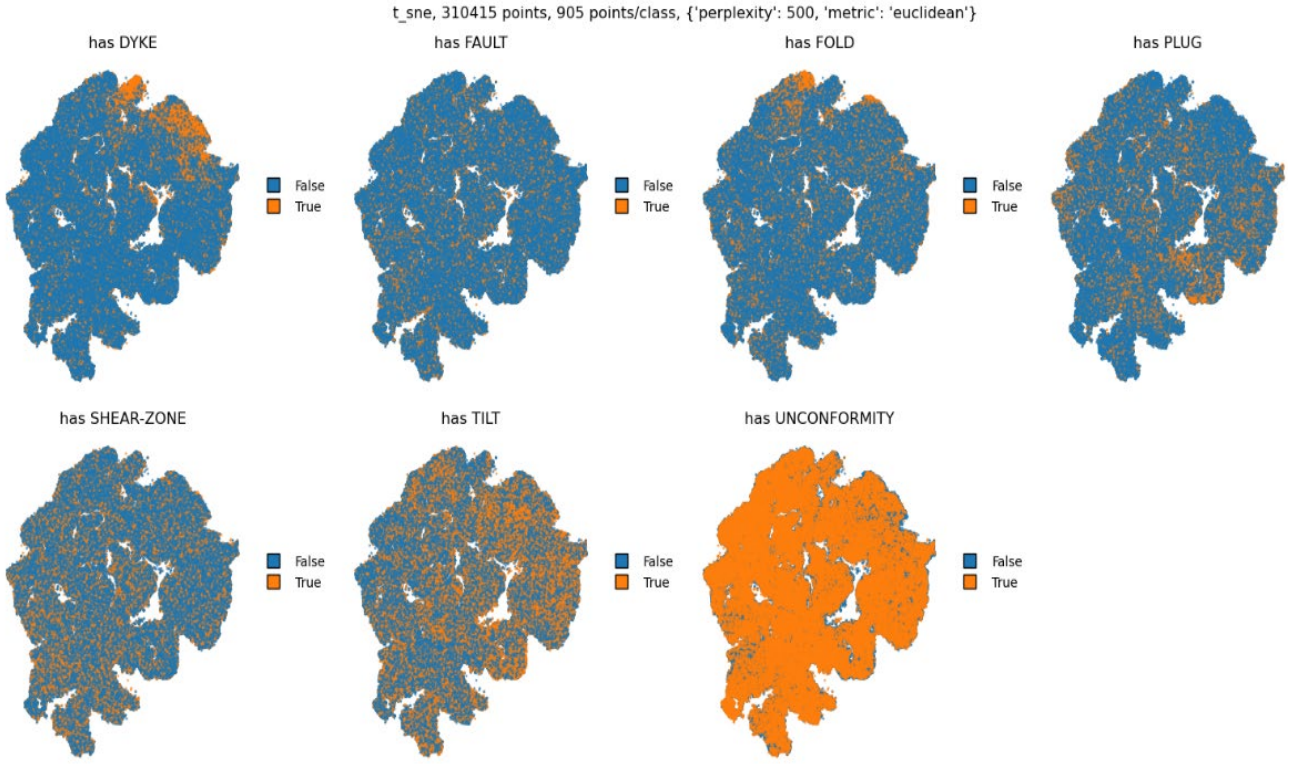
Visualized in only 2 dimensions...

Geological event history *[E1, E2, E3]* has event *X*?

PCA



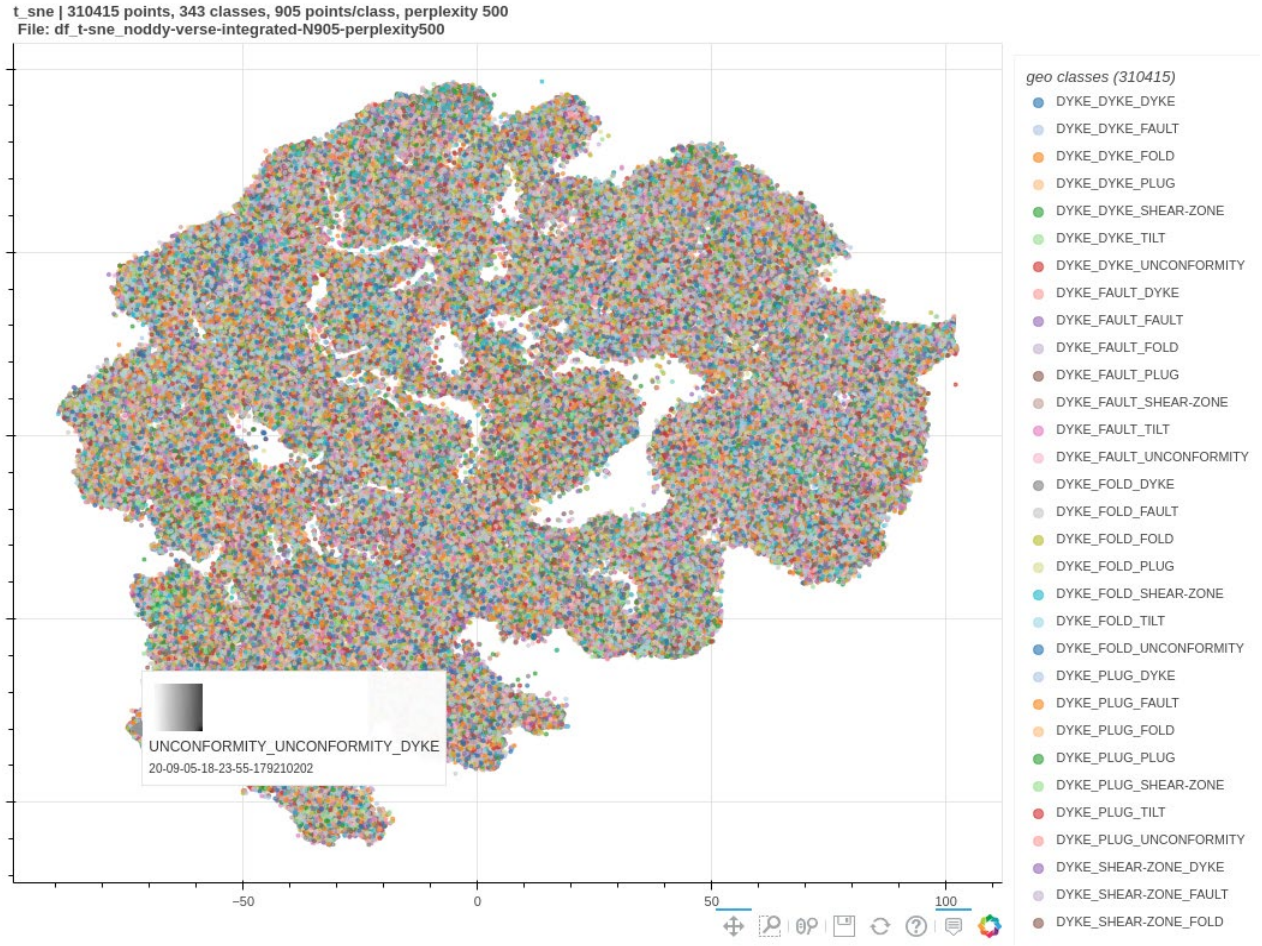
t-SNE



343 classes and 310K data points…

Encoded into only 13 dimensions...

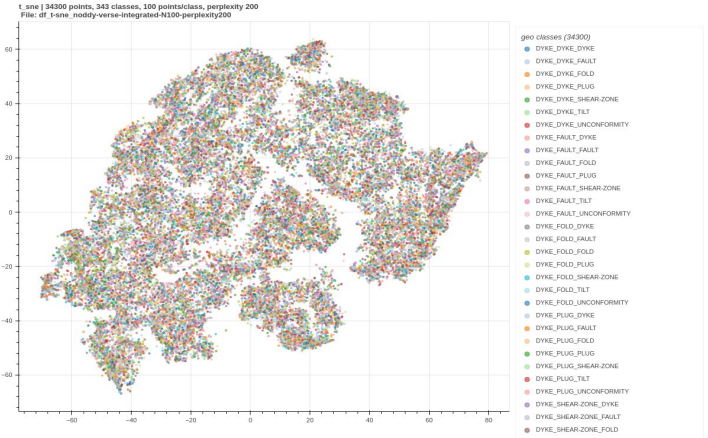
Visualized in only 2 dimensions...



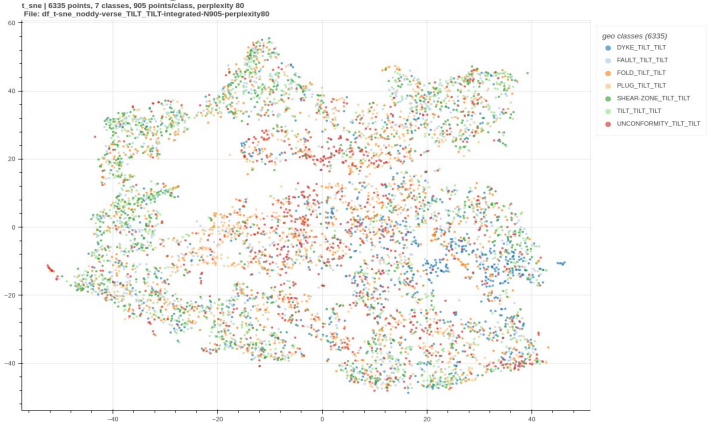
Hands-on

3-4 pm session

100 samples/class



Only [X, TILT, TILT]



Current and future work

Real-world data

Projection

NoddyVerse dataset

343 classes and 310K data points

“Maths”+Theory

Optimal (hyper) parameters, features and the curse of dimensionality

“Maths”+Theory

Auto-correlation function

- Haralick’s distance parameter optimal value ↔ scale of geological features

Adding/removing features to the embedding *vs* the curse/blessing of dimensionality

- Bespoke features
- Removing correlated features
- Other encoding (e.g., Fisher vector feature encoding)
- Data integration → Mag and Grav responses (26D space!)

Advanced t-SNE approaches

- Initializing via down sampling
- Perplexity annealing

Other non-liner projections?

- Uniform Manifold Approximation and Projection (**UMAP**), Multidimensional scaling (**MDS**), Self-organizing map (**SOM**)

REFERENCES

1) Jessell, M. et al. Into the Noddyverse: *a massive data store of 3D geological models for machine learning and inversion applications.* Earth Syst Sci Data 14, 381–392 (2021).

2) Guo, J. et al. *3D geological structure inversion from Noddy-generated magnetic data using deep learning methods*. Comput Geosci 149, 104701 (2021).

3) Haralick, R. M. *Statistical and structural approaches to texture*. P IEEE 67, 786–804 (1979).

4) Haralick, R. M., Shanmugam, K. & Dinstein, I. *Textural Features for Image Classification*. IEEE Transactions Syst Man Cybern SMC-3, 610–621 (1973).

5) Brynolfsson, P. et al. *Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters*. Sci Rep-uk 7, 4041 (2017).

6) Hall-Beyer, 2007. *GLCM Texture: A Tutorial* v. 1.0 through 2.7. http://hdl.handle.net/1880/51900

7) Van der Maaten, Laurens, and Geoffrey Hinton. *Visualizing data using t-SNE*. Journal of machine learning research 9, no. 11 (2008).

8) Kobak, D. & Berens, P. *The art of using t-SNE for single-cell transcriptomics*. Nat Commun 10, 5416 (2019).

9) Gorban, A. N., Tyukin, I. Y. *Blessing of dimensionality: mathematical foundations of the statistical physics of data.* Philosophical Transactions Royal Soc Math Phys Eng Sci 376, 20170237 (2018).

10) Köppen, M. *The curse of dimensionality*. in 5th online world conference on soft computing in industrial applications (WSC5) vol. 1 4–8 (2000).

THANK YOU!

Contact

Leonardo Portes

ll.portes@gmail.com

Haralick Features | Classifiers

SVM

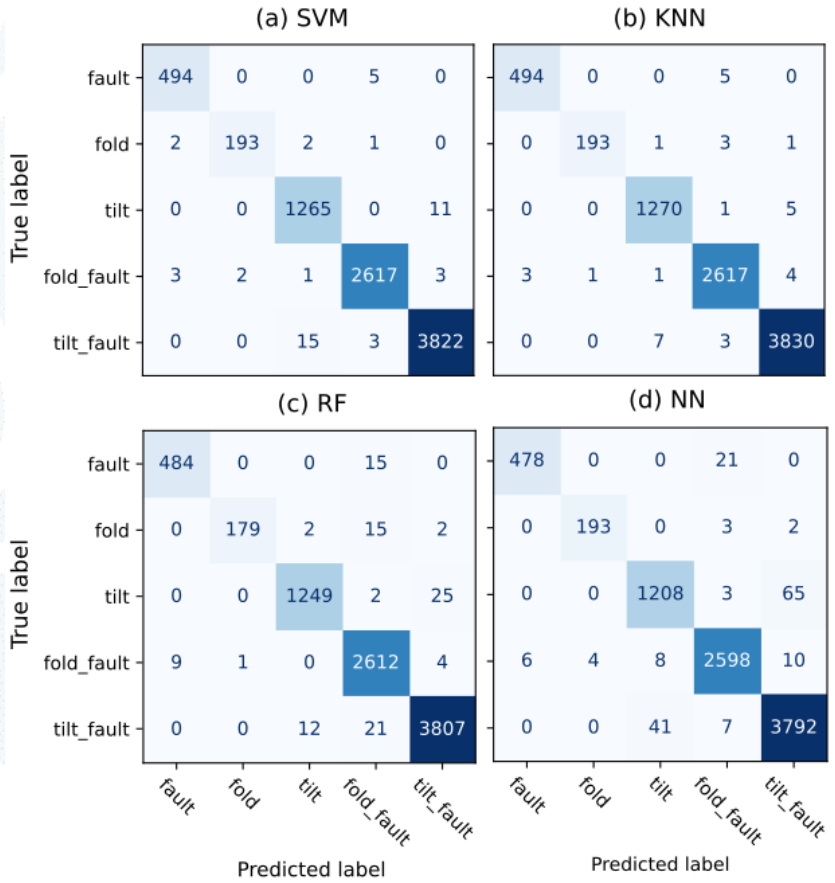
	precision	recall	f1-score	support
1_fault	1.00	0.99	0.99	499
2_fold	0.99	0.97	0.98	198
3_tilt	1.00	1.00	1.00	1276
4_fault_fold	1.00	1.00	1.00	2626
5_tilt_fault	1.00	1.00	1.00	3840
accuracy			1.00	8439
macro avg	1.00	0.99	0.99	8439
weighted avg	1.00	1.00	1.00	8439

KNN

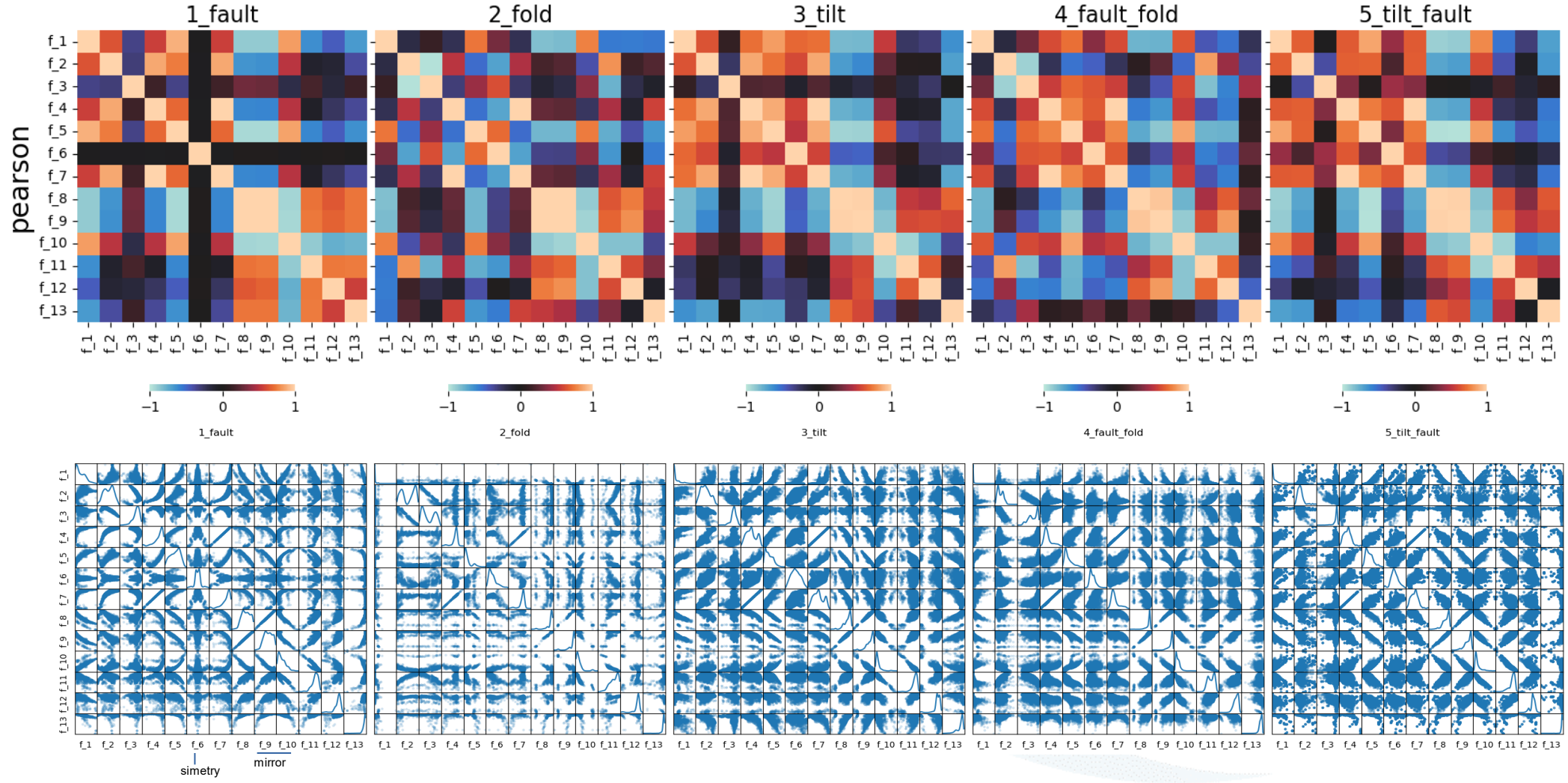
	precision	recall	f1-score	support
1_fault	0.99	0.99	0.99	499
2_fold	0.99	0.97	0.98	198
3_tilt	0.99	1.00	0.99	1276
4_fault_fold	1.00	1.00	1.00	2626
5_tilt_fault	1.00	1.00	1.00	3840
accuracy			1.00	8439
macro avg	0.99	0.99	0.99	8439
weighted avg	1.00	1.00	1.00	8439

RF

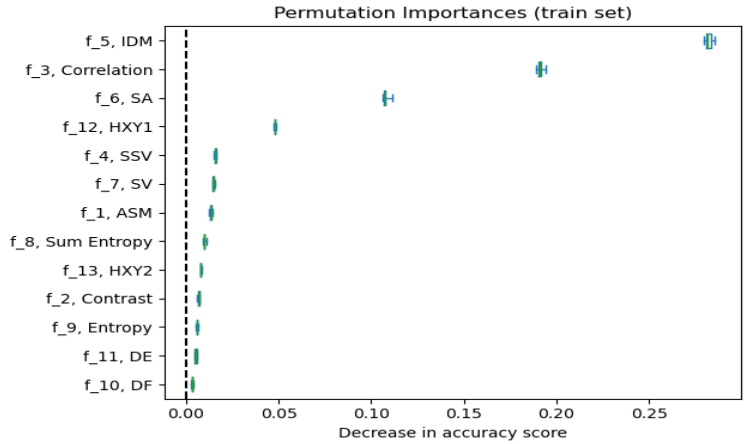
	precision	recall	f1-score	support
1_fault	0.96	0.95	0.96	499
2_fold	0.99	0.70	0.82	198
3_tilt	0.96	0.89	0.93	1276
4_fault_fold	0.96	0.99	0.97	2626
5_tilt_fault	0.96	0.98	0.97	3840
accuracy			0.96	8439
macro avg	0.97	0.90	0.93	8439
weighted avg	0.96	0.96	0.96	8439

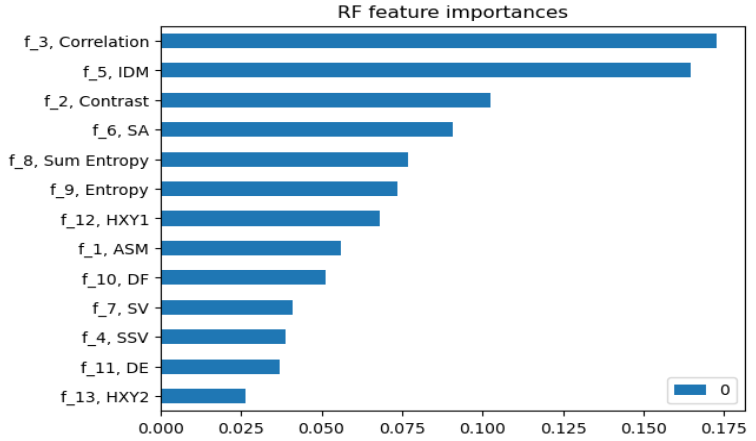


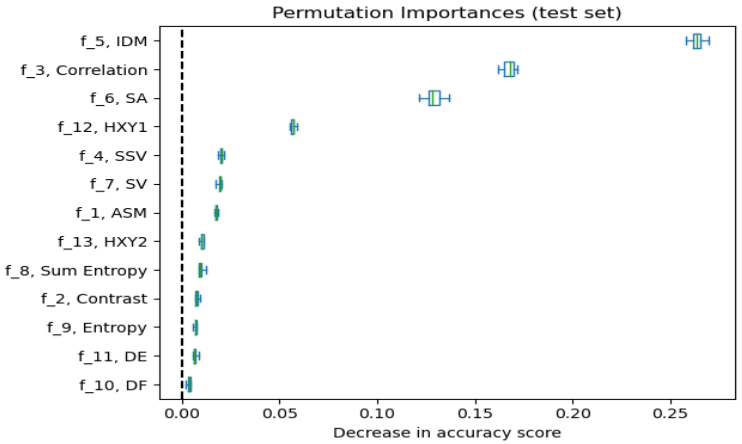
Haralick Features | Correlation



Haralick Features | Ranking







Pitfalls:

* imbalanced data
* importance related to classification task

notebooks/1.HaralickFeatures/2c.RF-training2022-08-17.ipynb#

The curse of dimensionality

Mario Köppen
Fraunhofer IPK Berlin
Pascalstr. 8-9, 10587 Berlin, Germany
E-Mail: *mario.koeppen@ipk.fhg.de*

Abstract

In this text, some question related to higher dimensional geometrical spaces will be discussed. The goal is to give the reader a feeling for geometric distortions related to the use of such spaces (e.g. as search spaces).

$V_{\text{hypersphere}}$ $\quad$ $V_{\text{hypercube}}$

$\frac{2r^d \pi^{d/2}}{d\,\Gamma(d/2)}$ $\quad$ $(2r)^d$

$$\frac{V_{\text{hypersphere}}}{V_{\text{hypercube}}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ as } d \to \infty$$

Blessing of dimensionality: mathematical foundations of the statistical physics of data

A. N. Gorban[1] and I. Y. Tyukin[1,2]

The concentrations of measure phenomena were discovered as the mathematical background to statistical mechanics at the end of the nineteenth/beginning of the twentieth century and have been explored in mathematics ever since. At the beginning of the twenty-first century, it became clear that the proper utilization of these phenomena in machine learning might transform the *curse of dimensionality* into the *blessing of dimensionality*. This paper summarizes recently discovered phenomena of measure concentration which drastically simplify some machine learning problems in high dimension, and allow us to correct legacy artificial intelligence systems. The classical