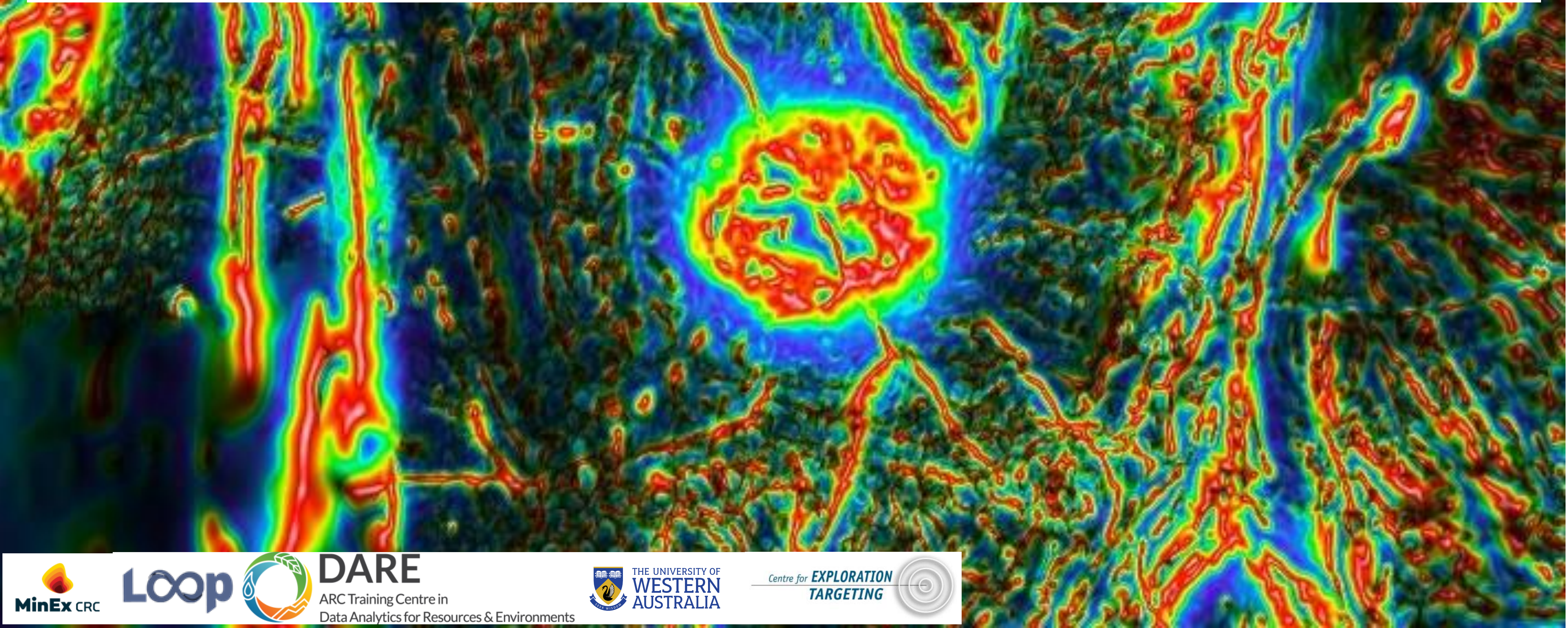
Combining feature engineering with non-linear projection to navigate geophysical datasets

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

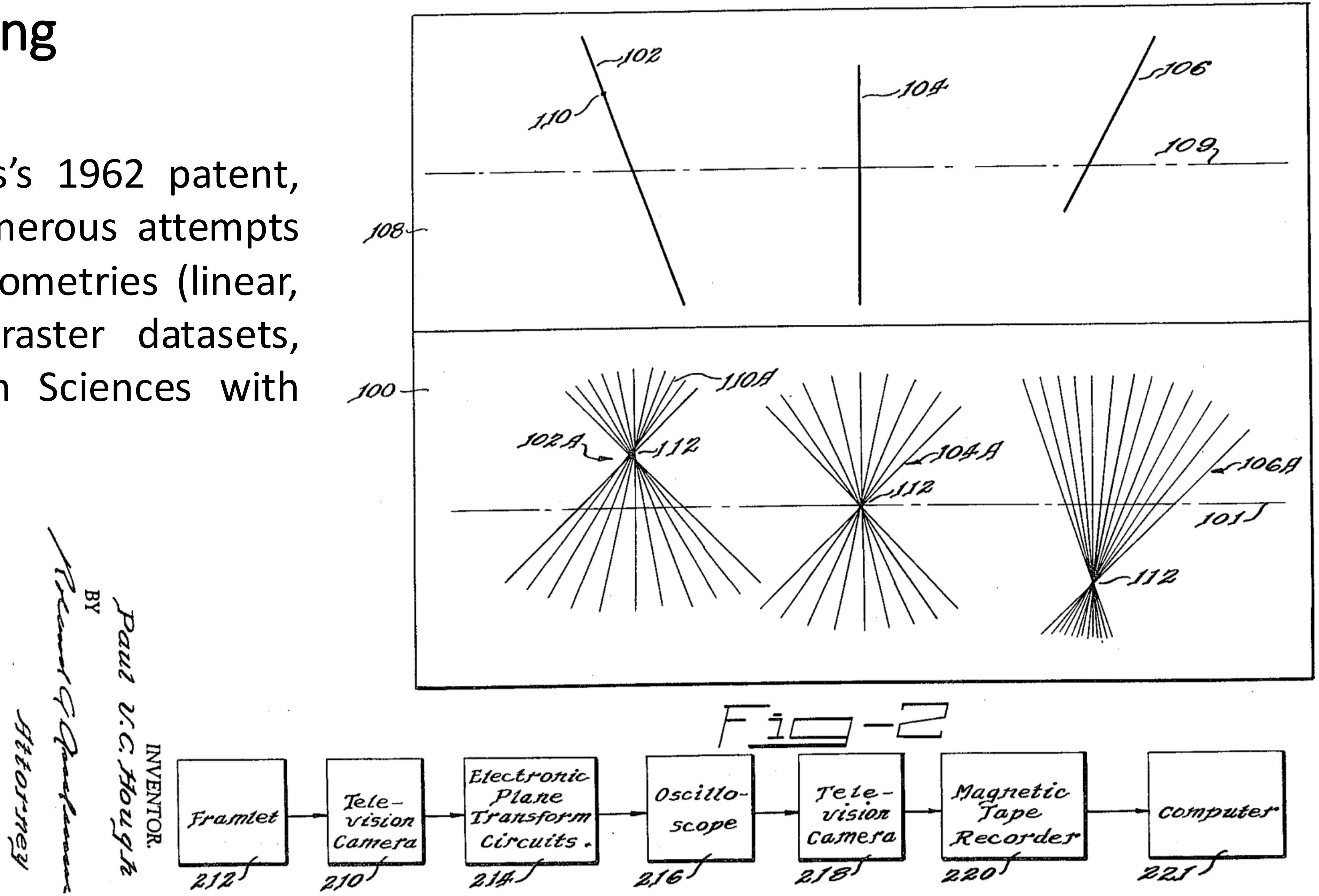
[1]UWA, [2]MinEx CRC, [3]ITTC DARE, [4]CSIRO



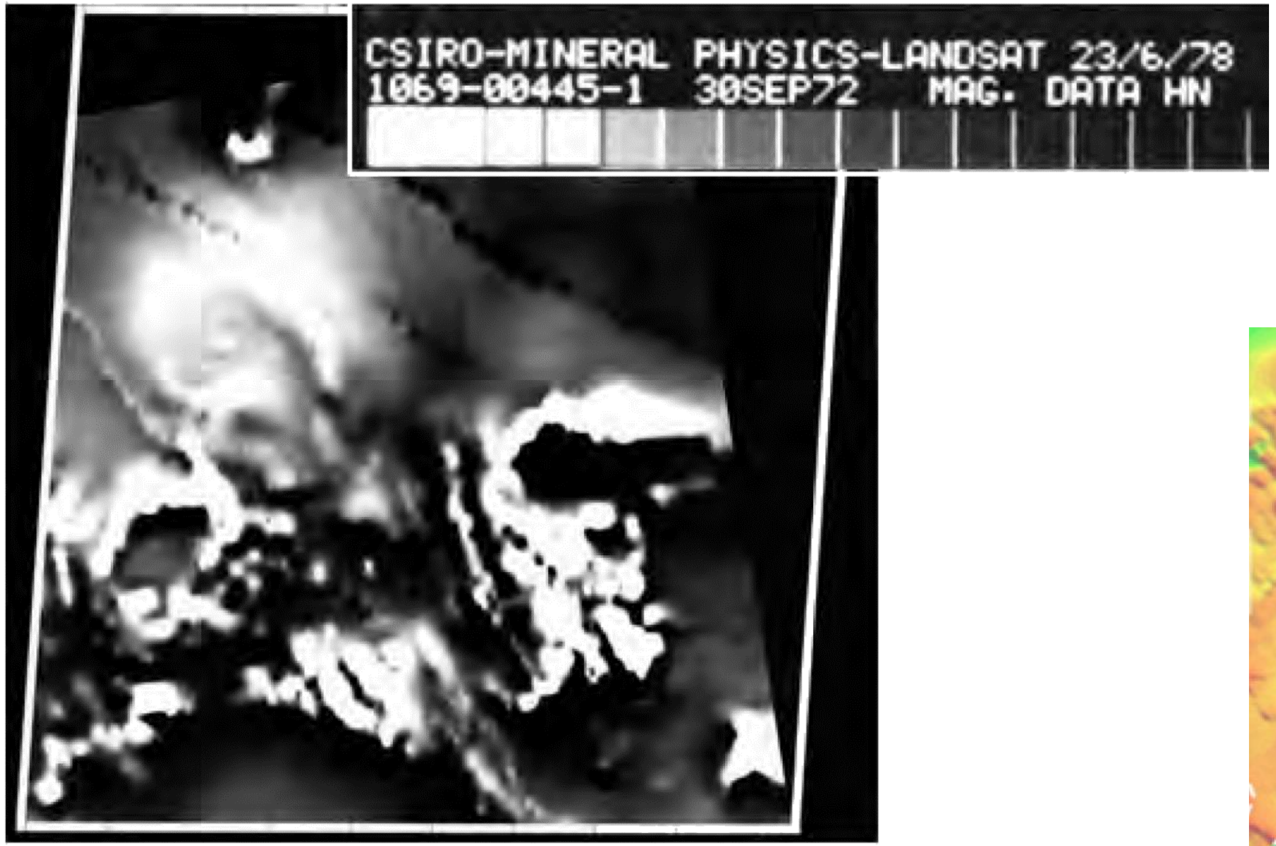
MinEx CRC
Loop
DARE
ARC Training Centre in
Data Analytics for Resources & Environments
THE UNIVERSITY OF WESTERN AUSTRALIA
Centre for EXPLORATION TARGETING

Template Matching

Starting with Houghs's 1962 patent, there have been numerous attempts to extract specific geometries (linear, circular etc.) from raster datasets, starting in the Earth Sciences with satellite data

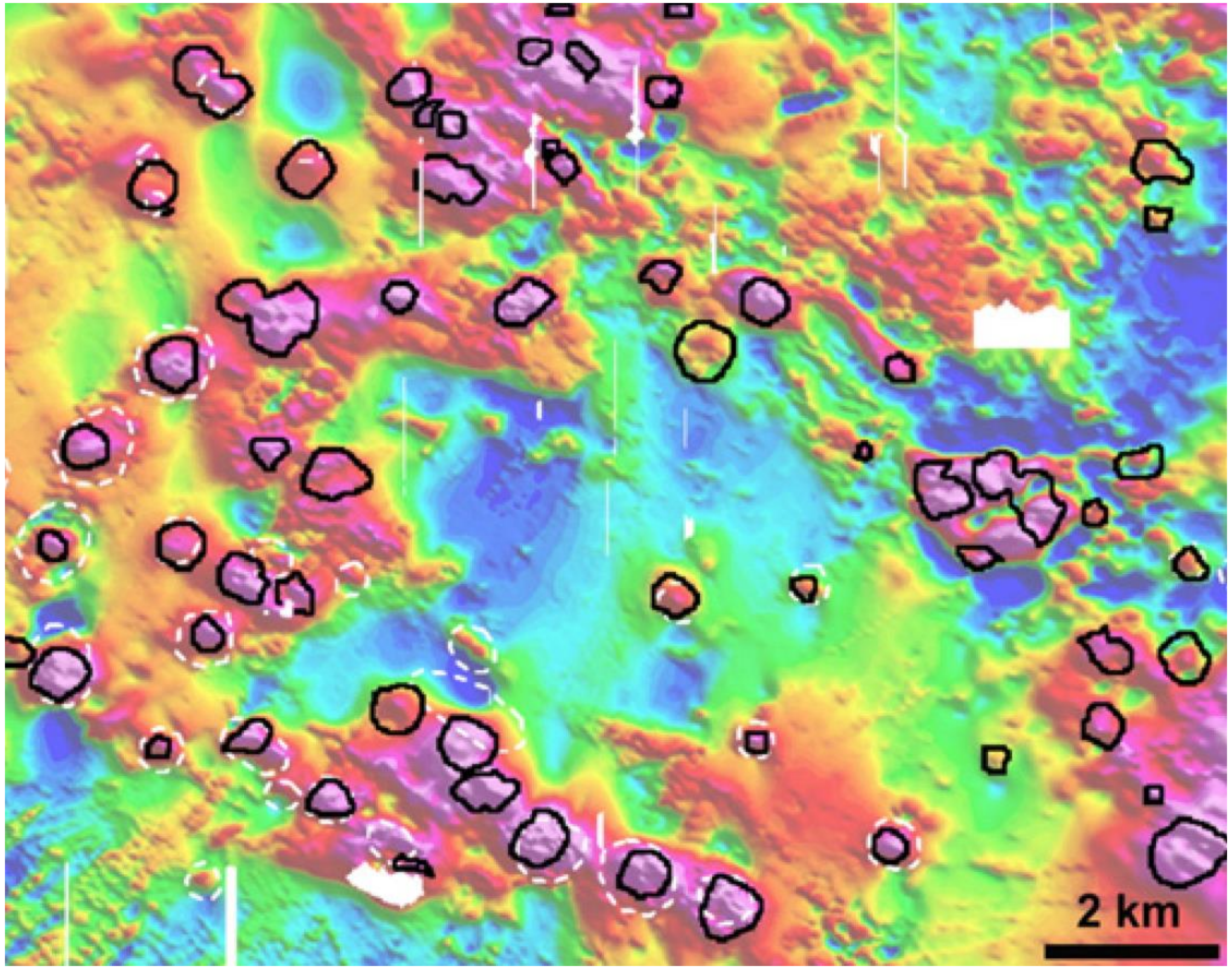


First Aeromagnetic Image - Pine Creek NT - 1978

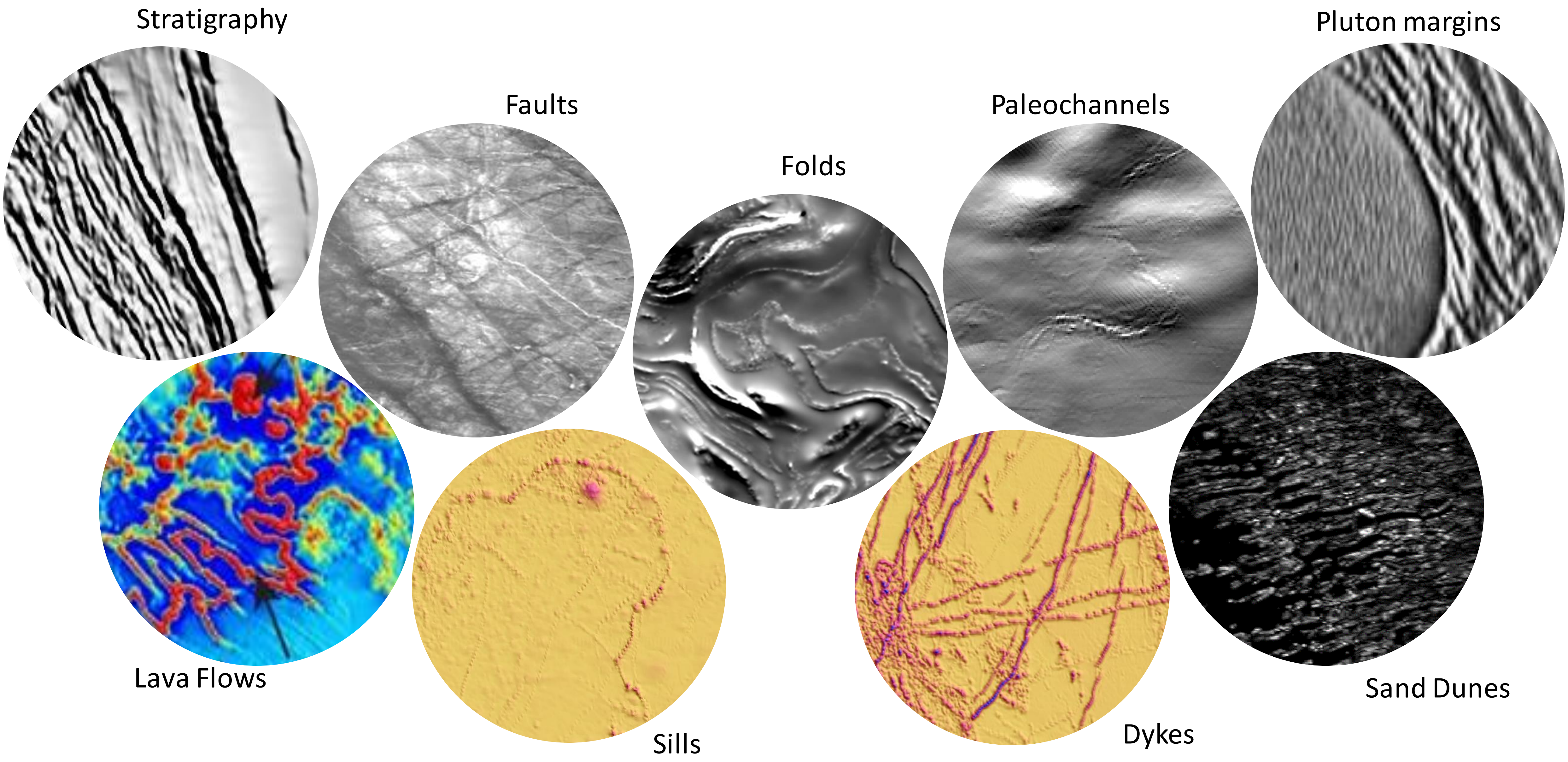


After the pioneering work of Huntington & Green in 1978 to produce the first gridded magnetic images, the same logic could then be applied to geophysical datasets.

Holden et al., 2011 Porphyry copper identification



Interpretation of single or multiple linear features



A problem…
and a vision

How to make sense of the **thousands of photos** we all have in our devices?

This is (some of) what I have on my phone





















Images from the web

My Little Big Planet

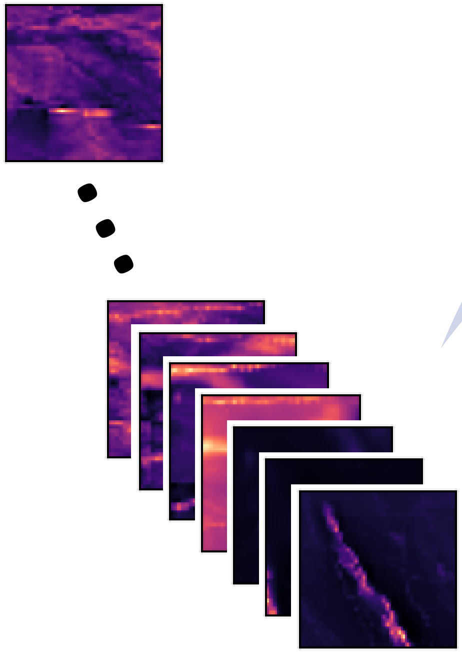
Images from the web

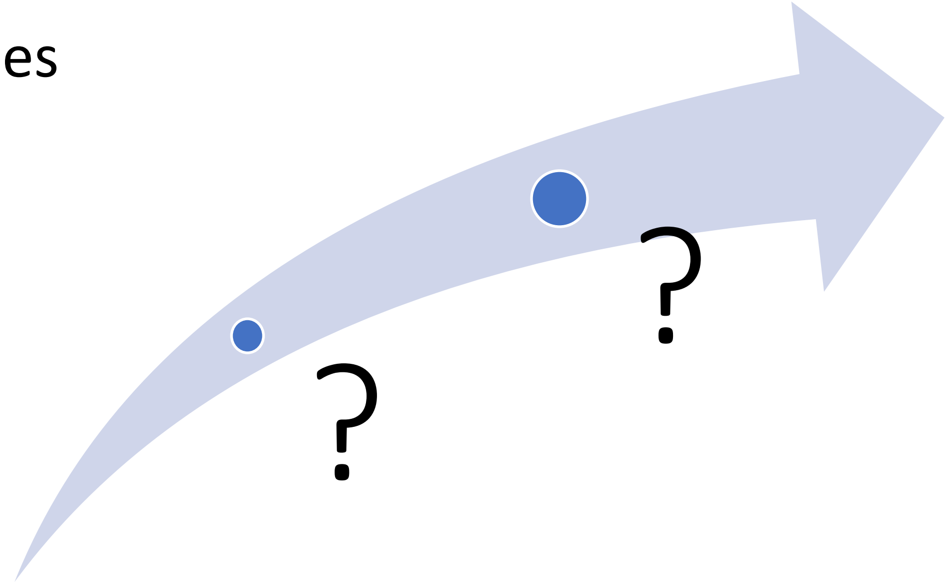

My Little Big Planet
"Peoples" Continent
"Animals" Interaction Continent
"Vehicles" Islands

My Little Big Planet
Hierarchical Organization!!!
"Peoples" Continent
Family Territory
Collective Sports Country
Individual Sports Country
"Animals" Interaction Continent
Farm Animals Republic
The Land of Dogs and Babies
Ocean Animals Nation
Anima Costume Island
"Vehicles" Islands
Driving
Being transported

Let's go back to Geosciences!

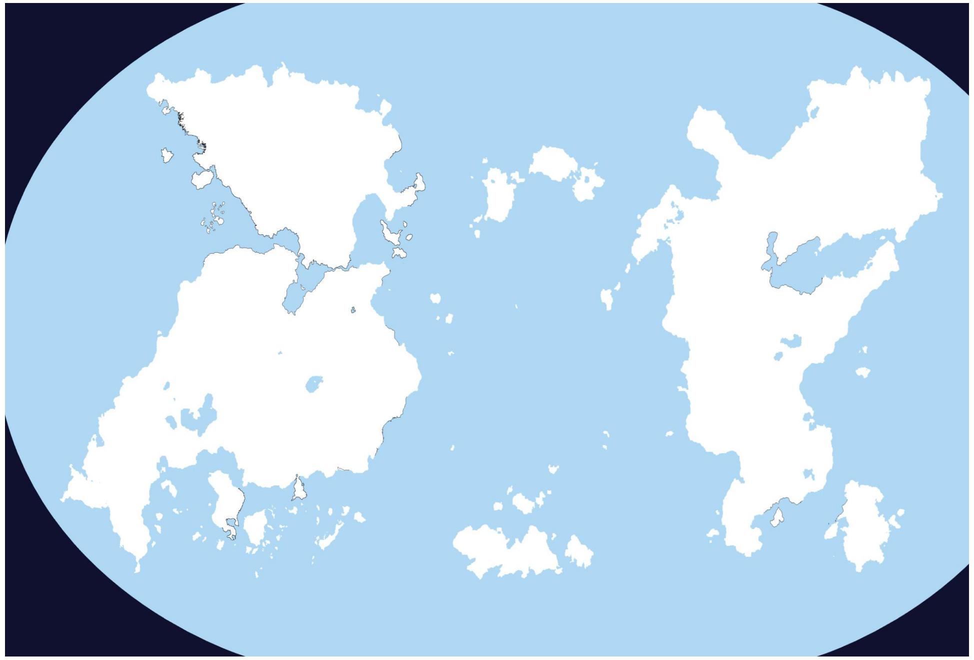
Many thousands of geophysical responses images



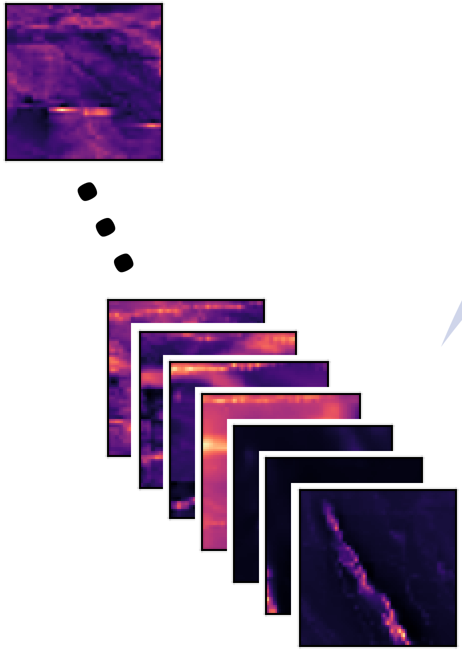


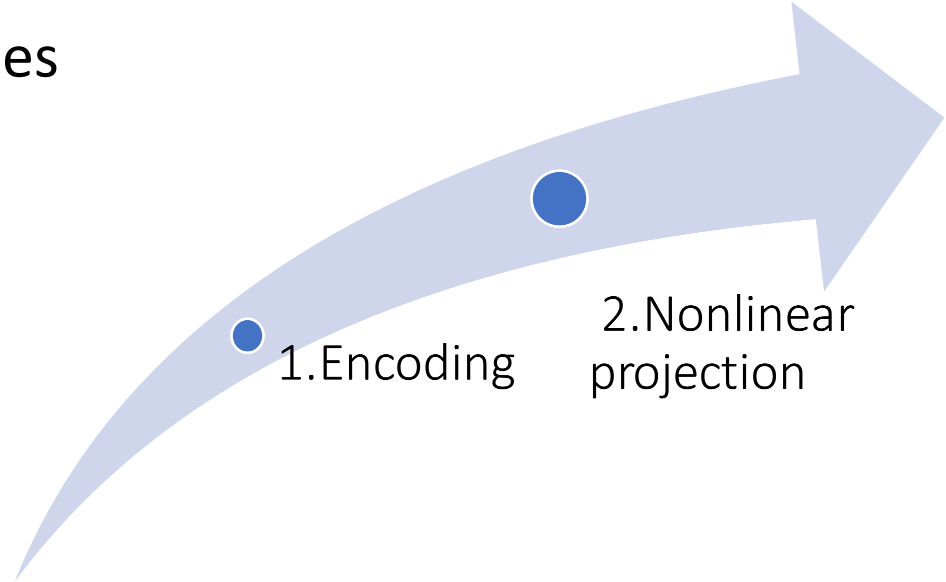
Requirements
- Speed
- Scalability
- Interpretability

Little Big Planet

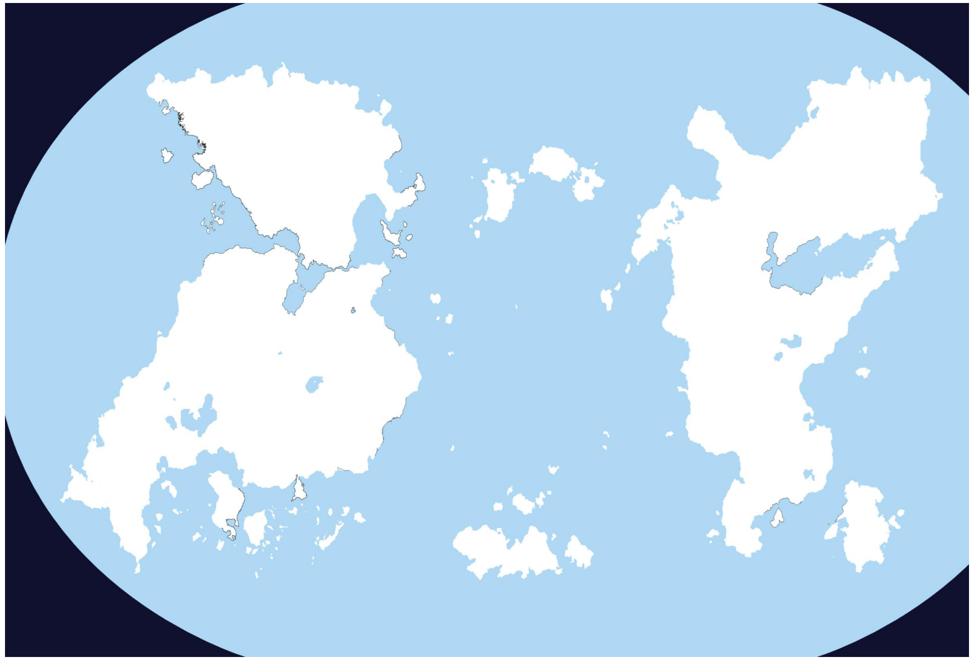


Many thousands of geophysical responses images





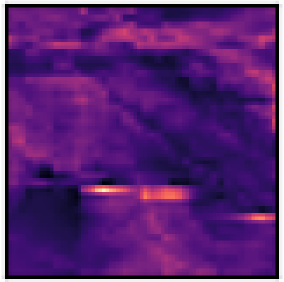
Little Big Planet



Requirements

- Speed
- Scalability
- Interpretability

1. Encoding | Haralick texture features



$$\mathbf{f} \doteq (f_1, f_2, \ldots, f_i, \ldots, f_{13}), \mathbf{f} \in \mathrm{Re}^{13}$$

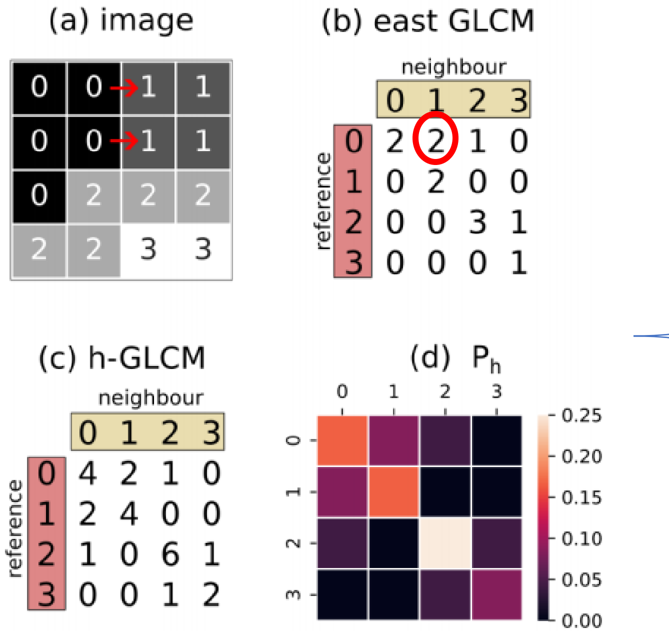


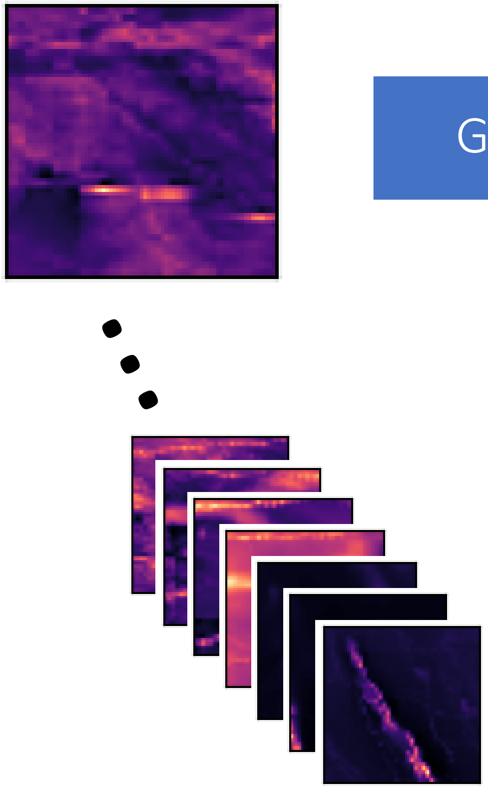
Figure 1: Illustrative example of the calculation of a GLCM for $(d, \theta) = (1, 0°)$.

- Repeat for $\theta \in \{0°, 45°, 90°, 135°\}$
- Calculate feature *fi* for each angle
- Get the mean

Example of a feature: **Entropy**

$$f_9 = -\sum_{i,j=1}^{n_g} p_{ij} \log(p_{ij})$$

1. Encoding | Haralick texture features

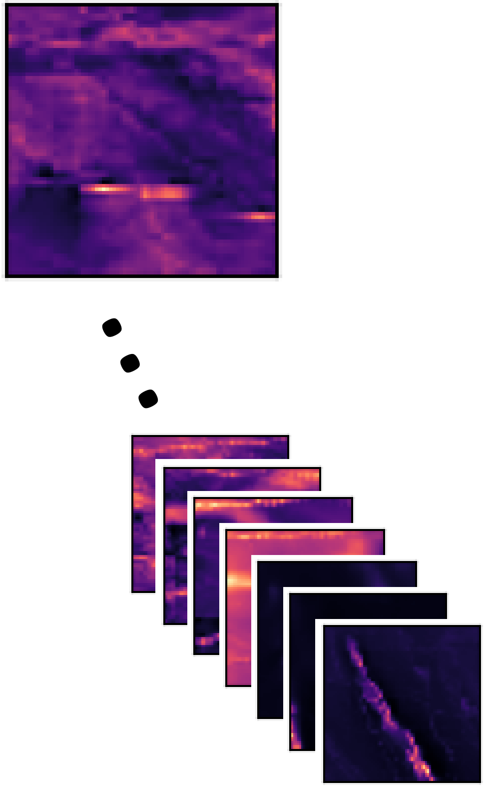


Each 1600-pixel image is represented as a single **point** in a 13D space!

Haralick Feature	Mathematical Expression	Notation
ASM	$\sum_{i,j=1}^{N} P(i,j)^2$	f_1
Contrast	$\sum_{i,j=1}^{N} (i-j)^2 P(i,j)$	f_2
Correlation	$\frac{\sum_{i,j=1}^{N} (i-\mu_X)(j-\mu_Y) P(i,j)}{\sigma_X \sigma_Y}$	f_3
Sum of Squares Variance	$\sum_{i,j=1}^{N} (i-\mu)^2 P(i,j)$	f_4
IDM	$\sum_{i,j=1}^{N} \frac{1}{1+(i-j)^2} P(i,j)$	f_5
Sum Average	$\sum_{k=2}^{2N} k P_{x+y}(k)$	f_6
Sum Variance	$\sum_{k=2}^{2N} (k - \text{Sum Average})^2 P_{x+y}(k)$	f_7
Sum Entropy	$-\sum_{k=2}^{2N} P_{x+y}(k) \log\left(P_{x+y}(k)\right)$	f_8
Entropy	$-\sum_{i,j=1}^{N} P(i,j) \log(P(i,j))$	f_9
Difference Variance	$\sum_{k=0}^{N-1} k^2 P_{x-y}(k)$	f_{10}
Difference Entropy	$-\sum_{k=0}^{N-1} P_{x-y}(k) \log\left(P_{x-y}(k)\right)$	f_{11}
IMC_1	$\frac{HXY - HXY1}{\max(HX, HY)}$	f_{12}
IMC_2	$\sqrt{1-\exp(-2(HXY2 - HXY))}$	f_{13}

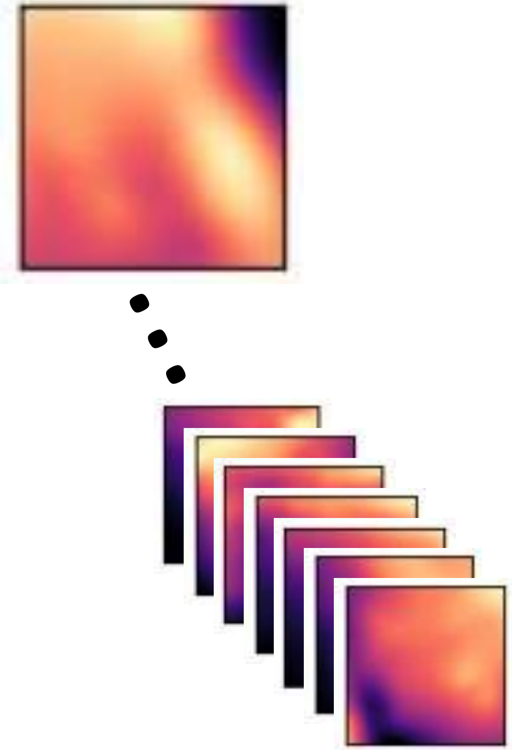
1. Encoding | play this game for each type of geophysical data

Magnetics



$$\mathbf{f_m} \doteq (f_{m1}, \ldots, f_{m13})$$

Gravity



$$\mathbf{f_g} \doteq (f_{g1}, \ldots, f_{g13})$$

1. Encoding | **(BONUS!) make a joint analysis!**

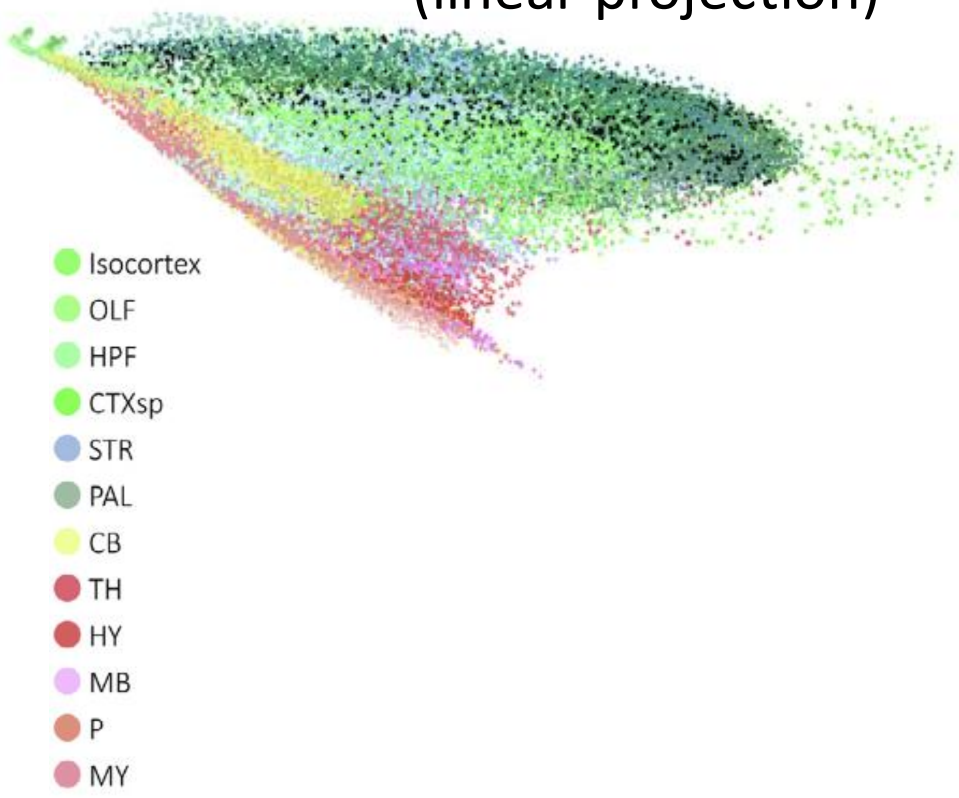
$$\mathbf{f_{mg}} \doteq (f_{m1}, \ldots, f_{m13}, f_{g1}, \ldots, f_{g13}), \mathbf{f_{mg}} \in \mathbb{R}^{26}$$

2. Nonlinear projection | t-SNE (t-distributed Stochastic Neighbor Embedding)

c.f. Horrocks et al. 2019

Principal Component Analysis (PCA)

(linear projection)



Coronal mouse brain transcriptome similarities [...] colored by anatomical region labels

A. Mahfouz et al. (2015). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings

2. Nonlinear projection | t-SNE

PCA

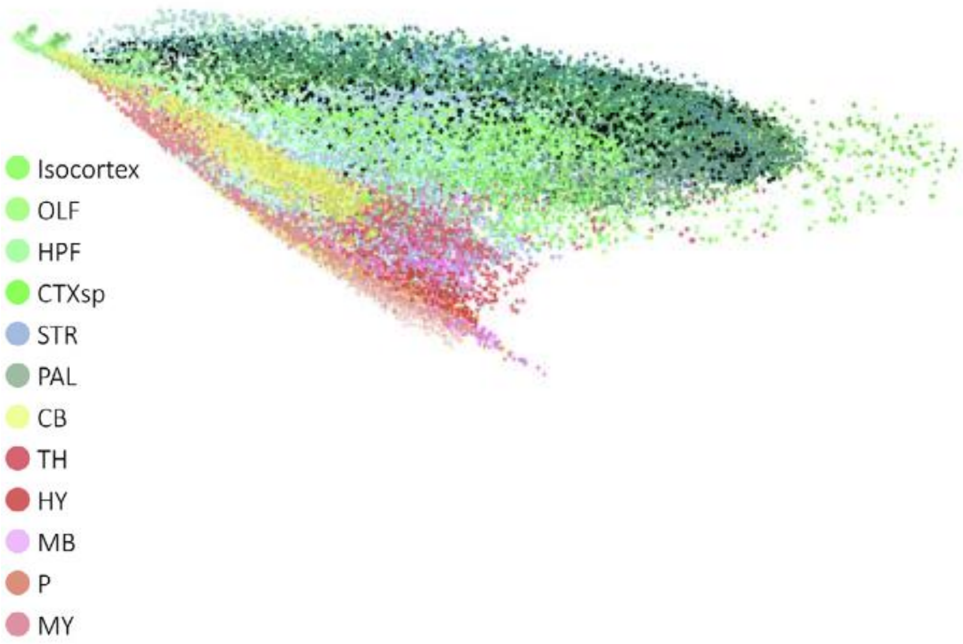
(linear projection)

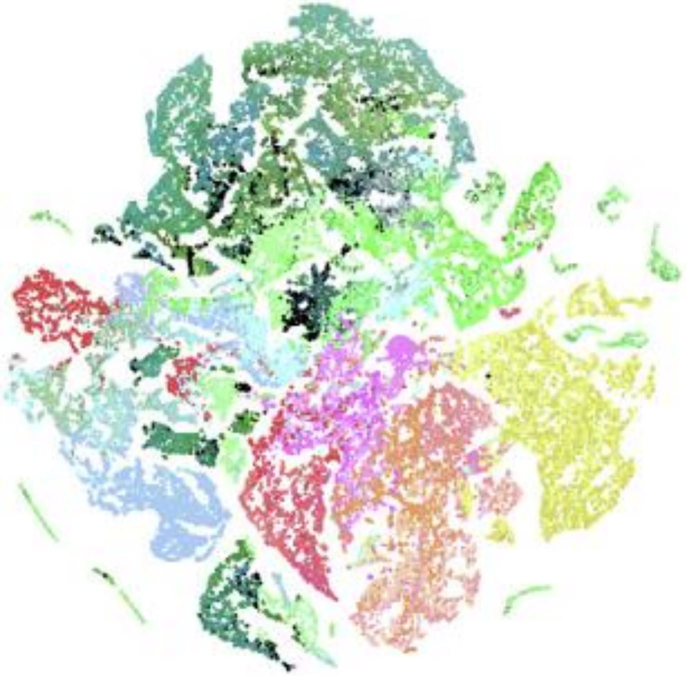
t-SNE

(nonlinear projection)

Developed by van der Maaten & Hinton (2008)

Coronal mouse brain transcriptome similarities [...] colored by anatomical region labels





A. Mahfouz et al. (2015). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings

2. Nonlinear projection | t-SNE

Cast as "probabilities"

(over a probability distribution function)

Pairwise **distances** in high dimensional space (ORIGINAL DATA)

Gaussian

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)}$$

Student-T

$$q_{i|j} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}}$$

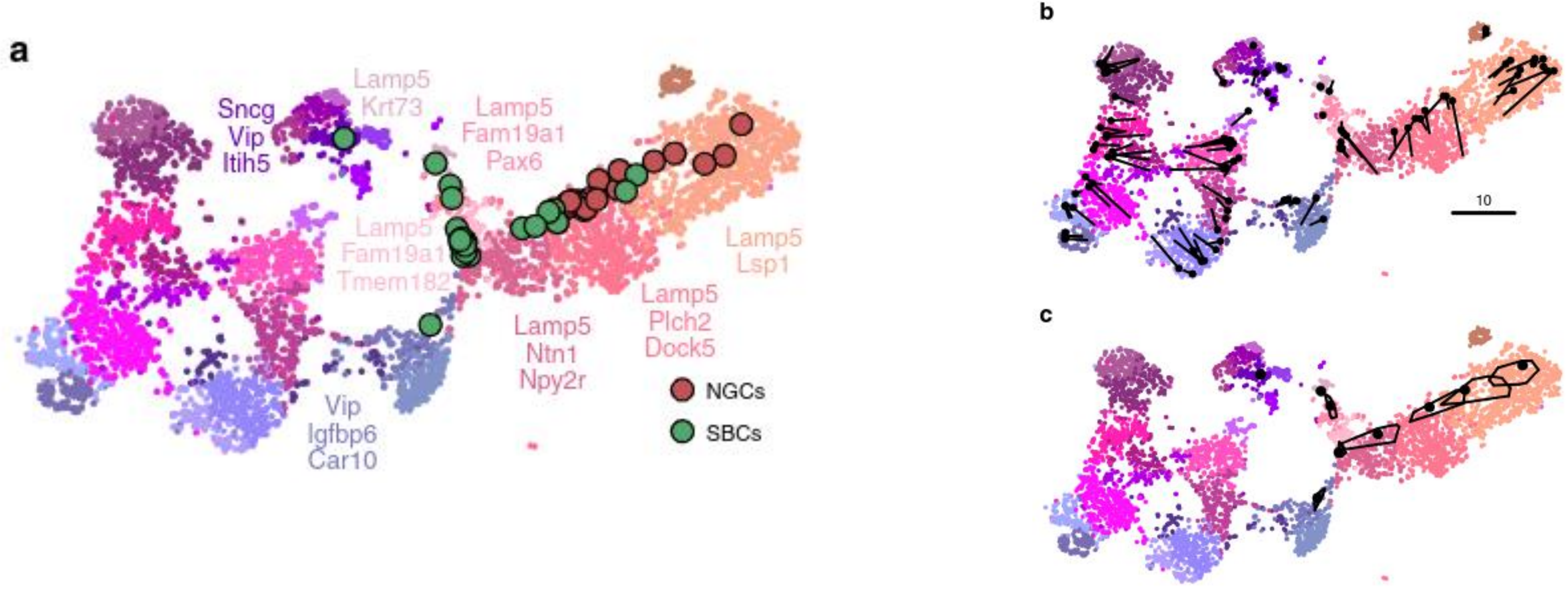
Pairwise **distances** in 2D (THE Projection!)

KL divergence (Kullback-Leibler)

Move points y_i in the direction of lower KL!

2. Nonlinear projection | t-SNE

Visualizations as **maps of meaning**!

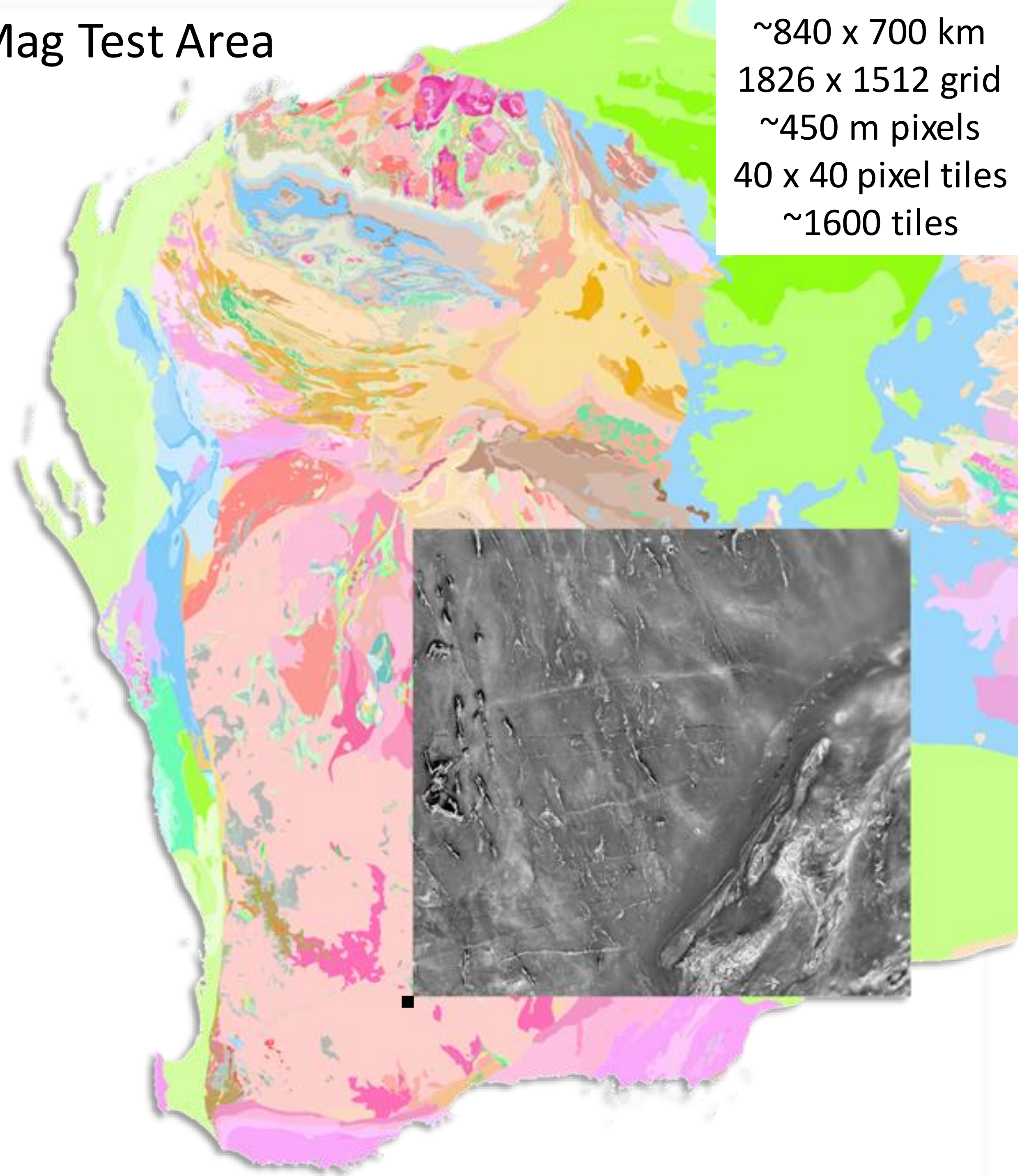


D. Kobak, P. Berens (2019). The art of using t-SNE for single-cell transcriptomics

t-SNE Image Atlas

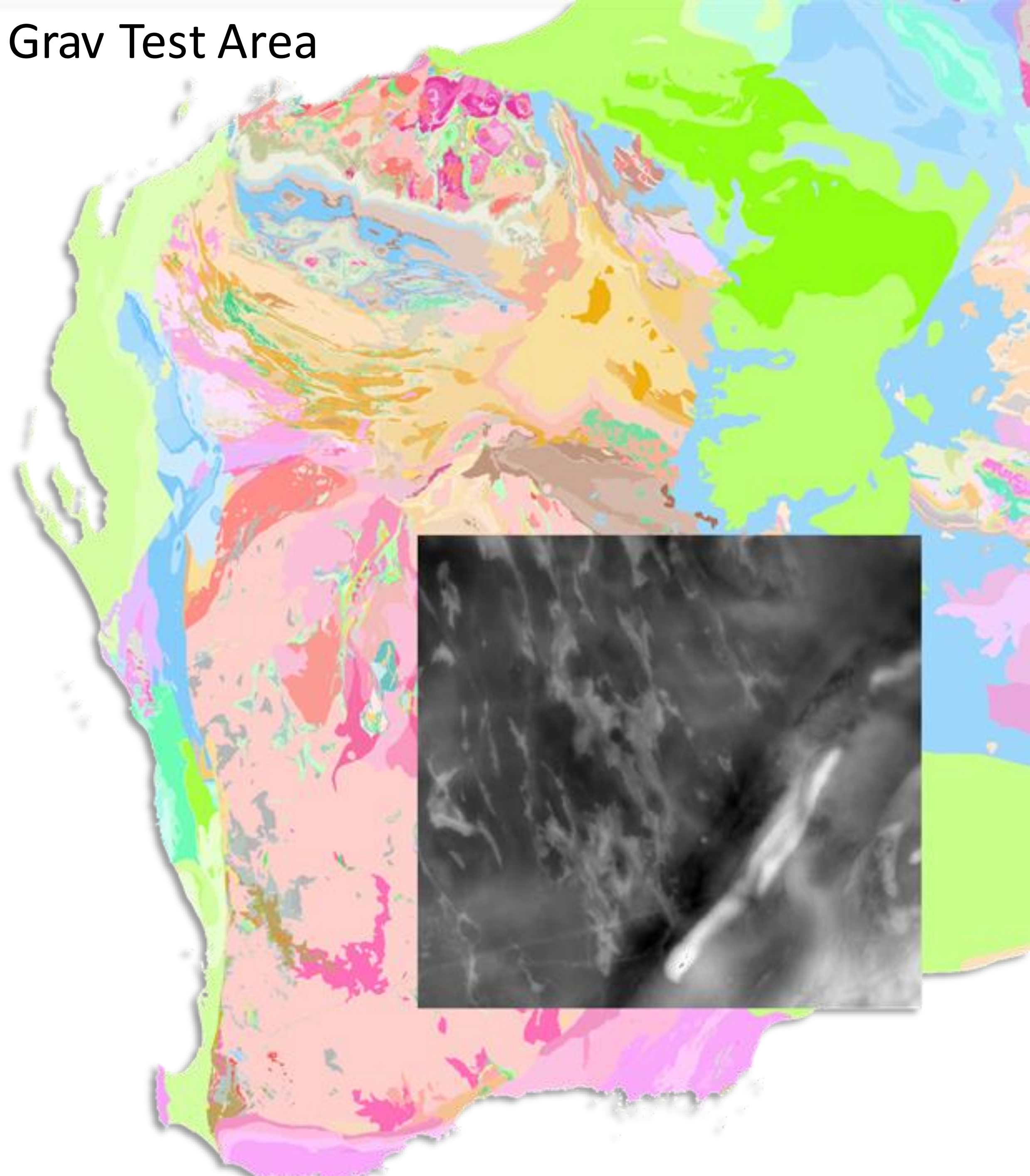
applied to geophysical data from Western Australia

Mag Test Area



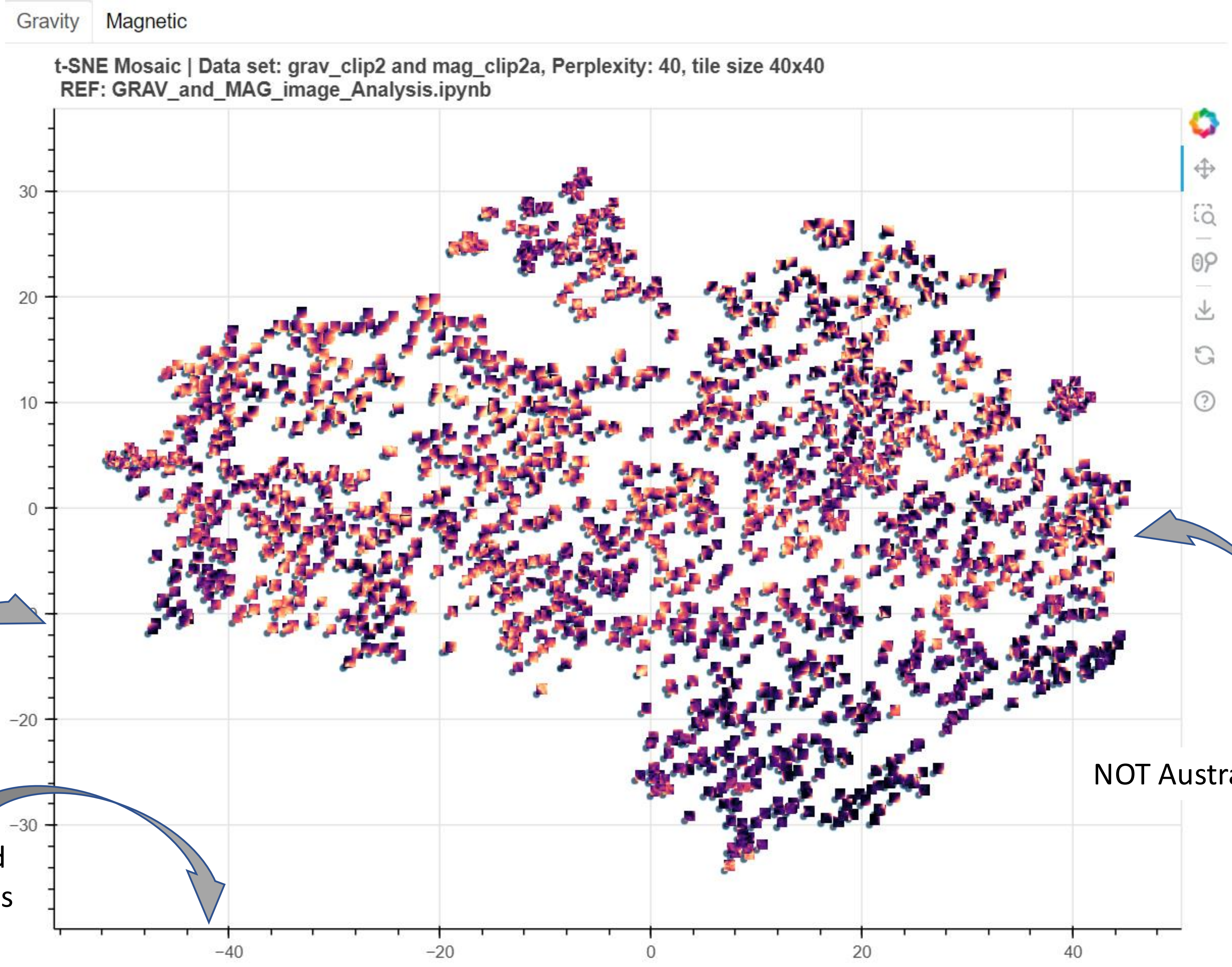
~840 x 700 km
1826 x 1512 grid
~450 m pixels
40 x 40 pixel tiles
~1600 tiles

Grav Test Area



Grav

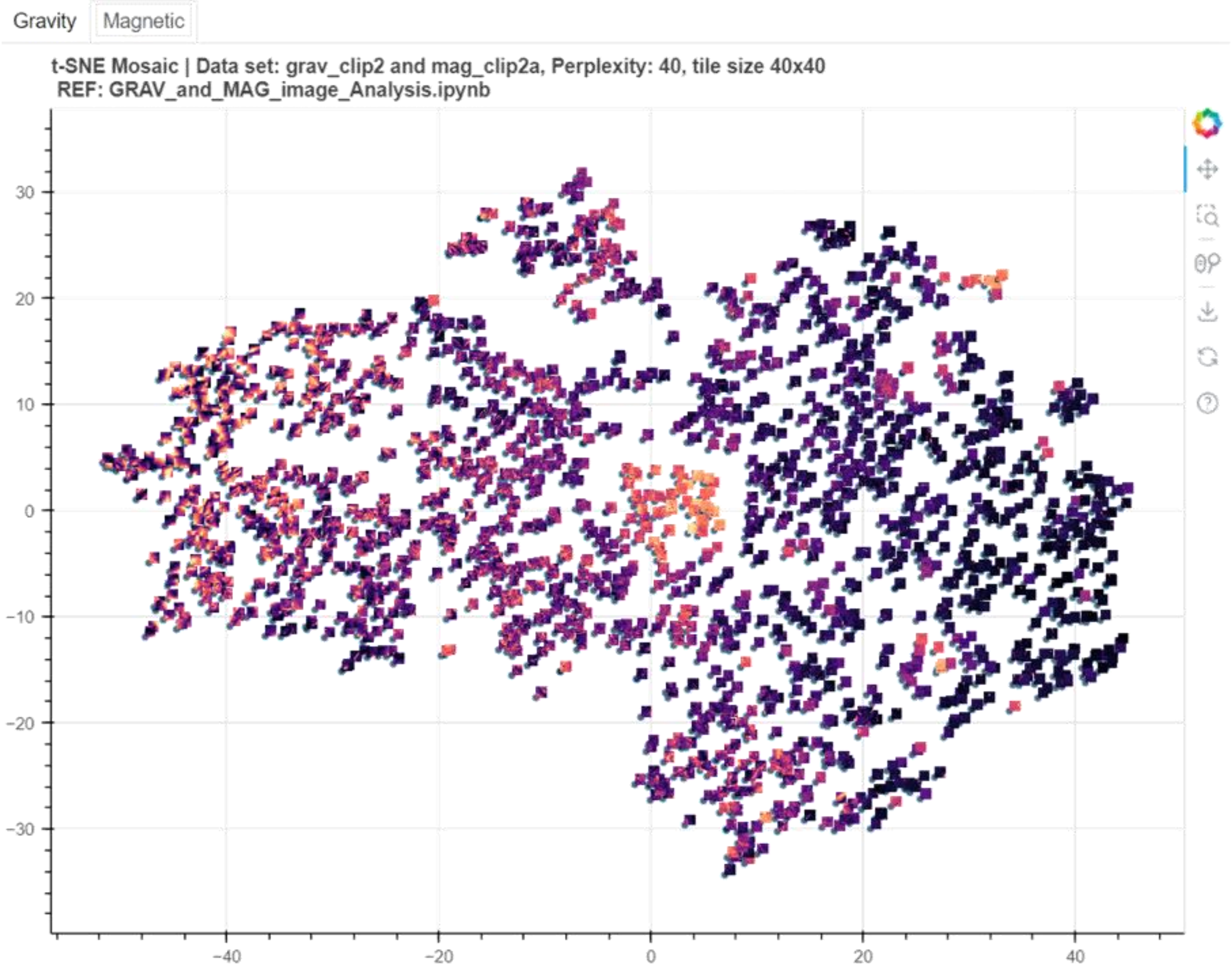
~1600 tiles



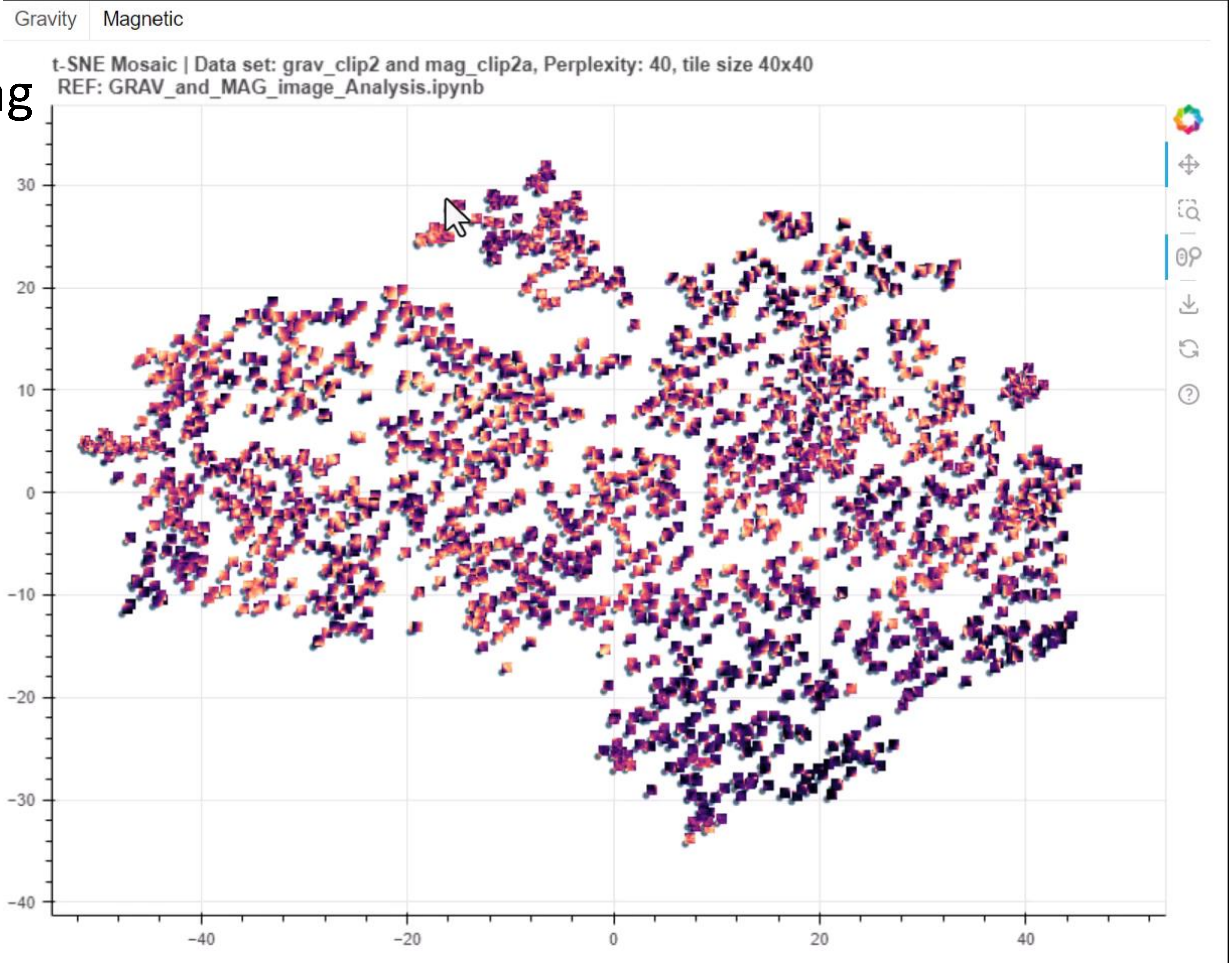
NOT real-world geographic axes

NOT Australia!

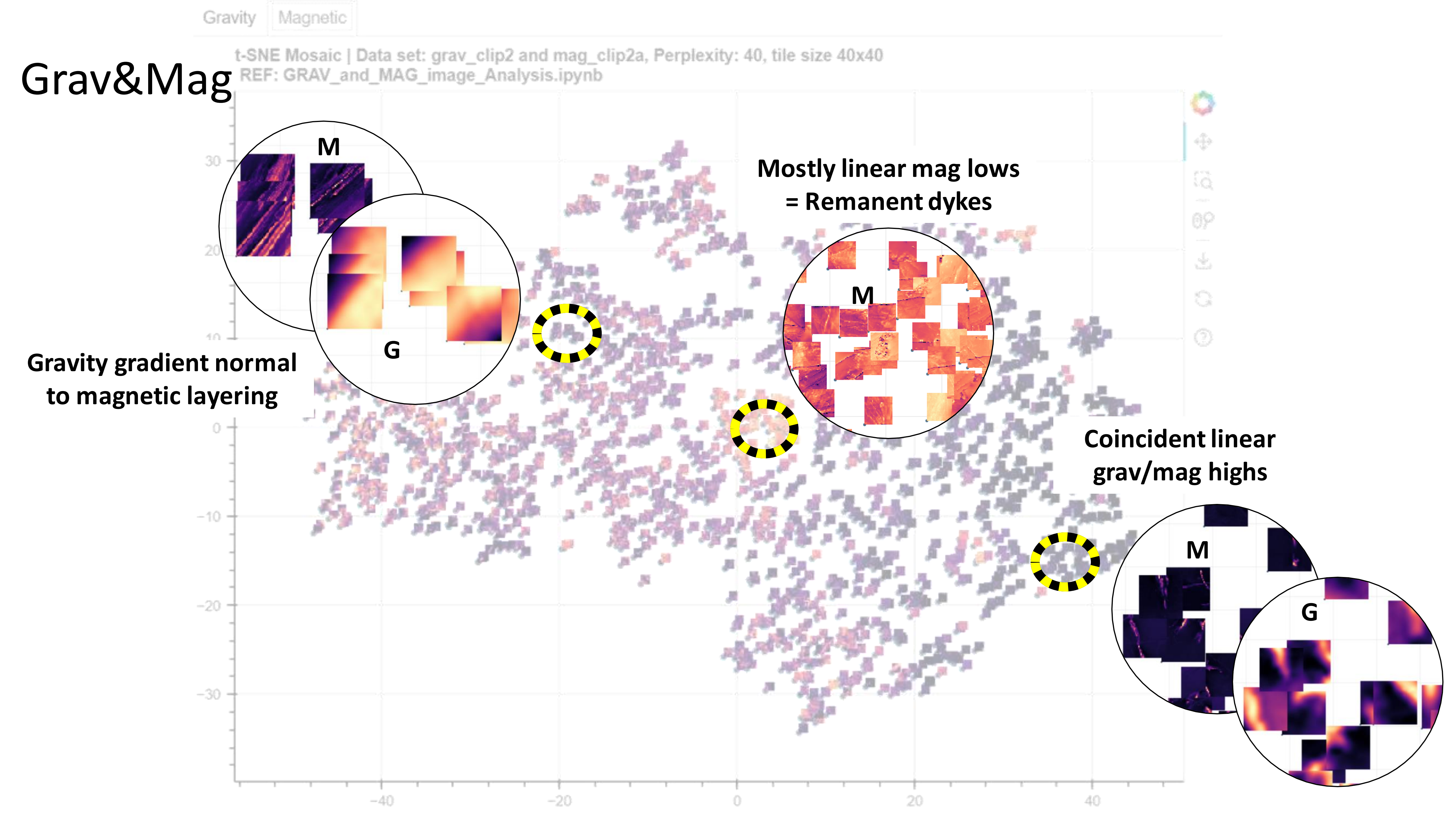
Mag


Gravity
Magnetic
t-SNE Mosaic | Data set: grav_clip2 and mag_clip2a, Perplexity: 40, tile size 40x40
REF: GRAV_and_MAG_image_Analysis.ipynb
30
20
10
0
−10
−20
−30
−40
−20
0
20
40

Grav&Mag


Gravity
Magnetic
t-SNE Mosaic | Data set: grav_clip2 and mag_clip2a, Perplexity: 40, tile size 40x40
REF: GRAV_and_MAG_image_Analysis.ipynb
30
20
10
0
-10
-20
-30
-40
-40
-20
0
20
40

Grav&Mag


Gravity
Magnetic
t-SNE Mosaic | Data set: grav_clip2 and mag_clip2a, Perplexity: 40, tile size 40x40
REF: GRAV_and_MAG_image_Analysis.ipynb
M
G
Gravity gradient normal to magnetic layering
Mostly linear mag lows = Remanent dykes
M
Coincident linear grav/mag highs
M
G
30
20
10
0
−10
−20
−30
−40
−20
0
20
40

Comments

- Allows us to separate different grav, mag and grav/mag anomaly shapes
- Data-centric, not dependant on prior assumptions about geometry
- Can easily be applied to any type(s) of gridded data
- Could replace Haralick Features with a Convolutional Neural Network
- Obviously, could include consideration of scale:
    - Size of image tile [40x40]
    - Offset neighbour analysis in Haralick feature extraction [1]
    - Sliding window offset instead of tiled patches [40]
- Could plot clusters back in real-world positions…
- Could develop web version of tool so anyone can use it?

2. Nonlinear projection | t-SNE