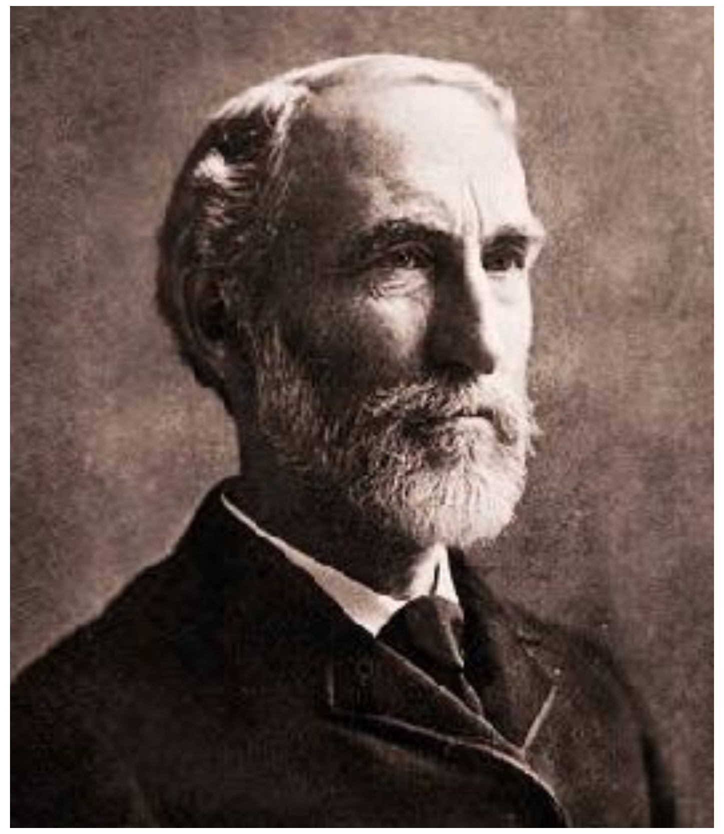
data science and science with data

Brice Ménard

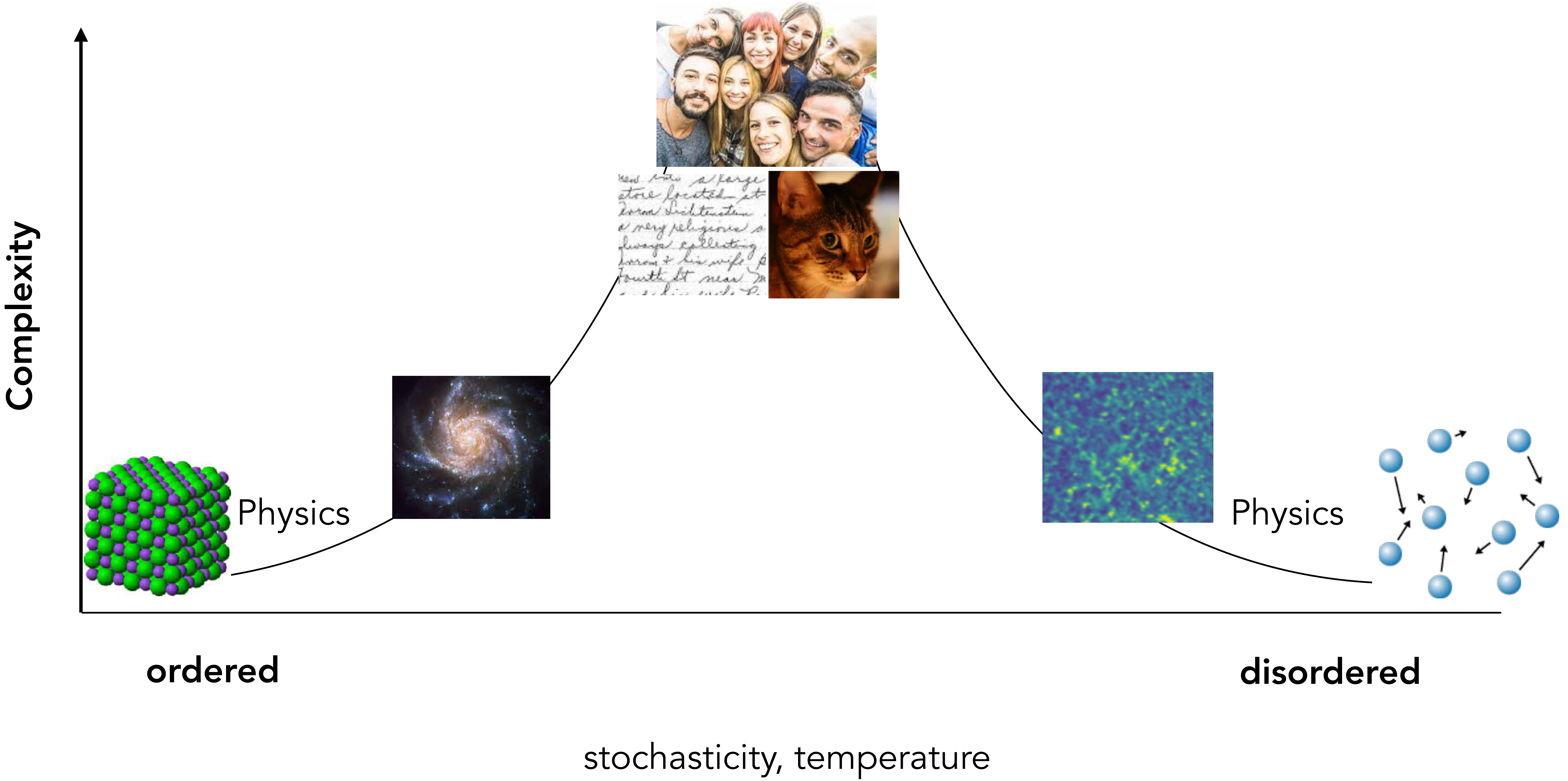
Johns Hopkins University
Ecole Normale Supérieure



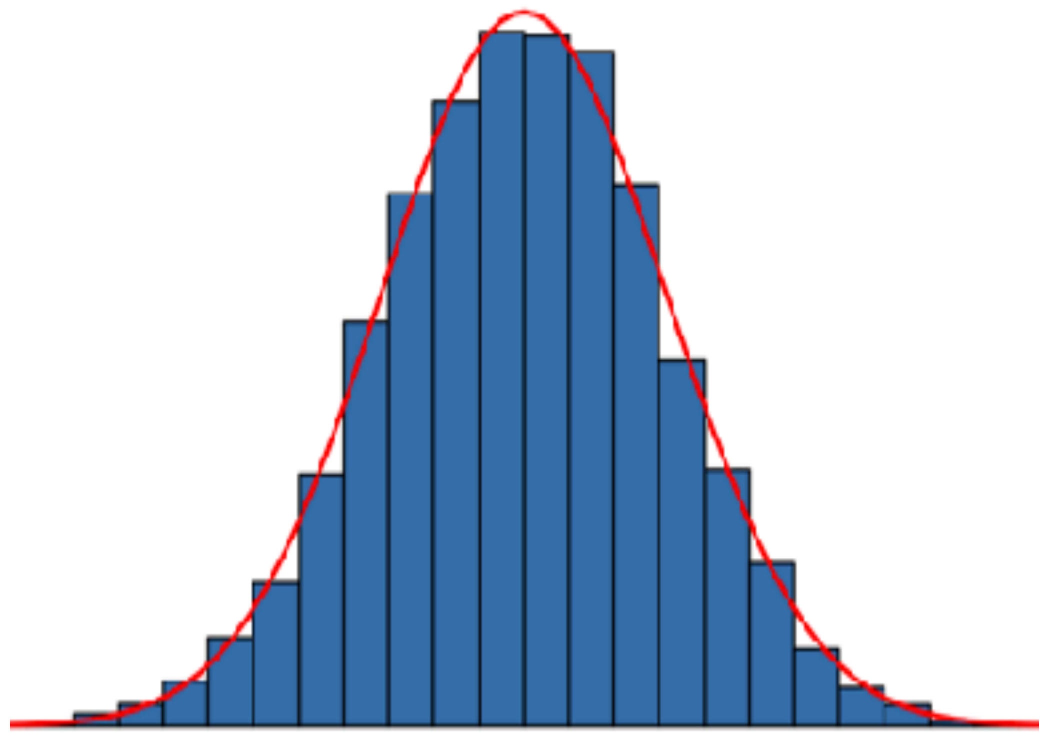
"One of the principal objects of theoretical research is to find the point of view from which the subject appears in the greatest simplicity."

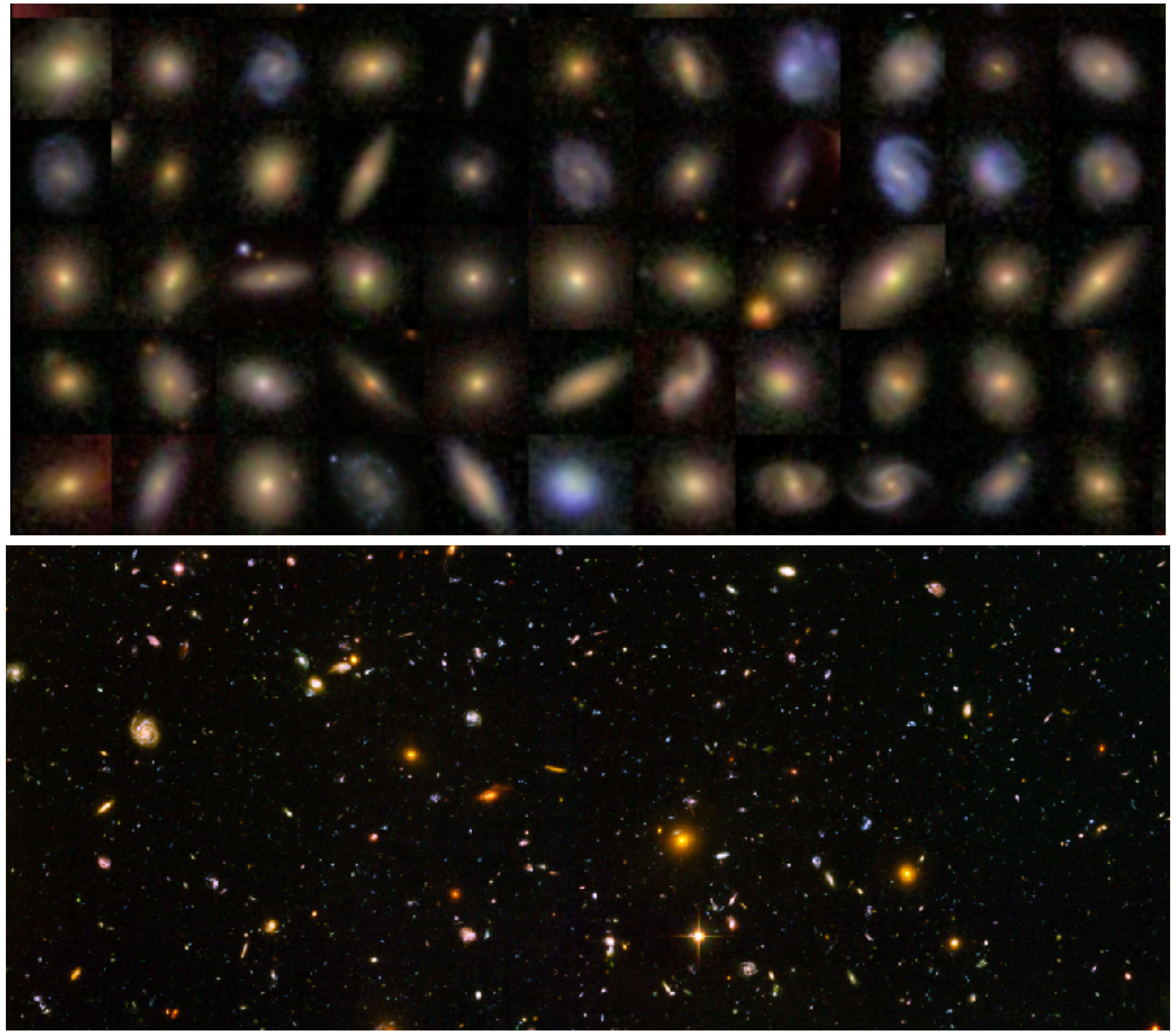
(Gibbs, 1881)

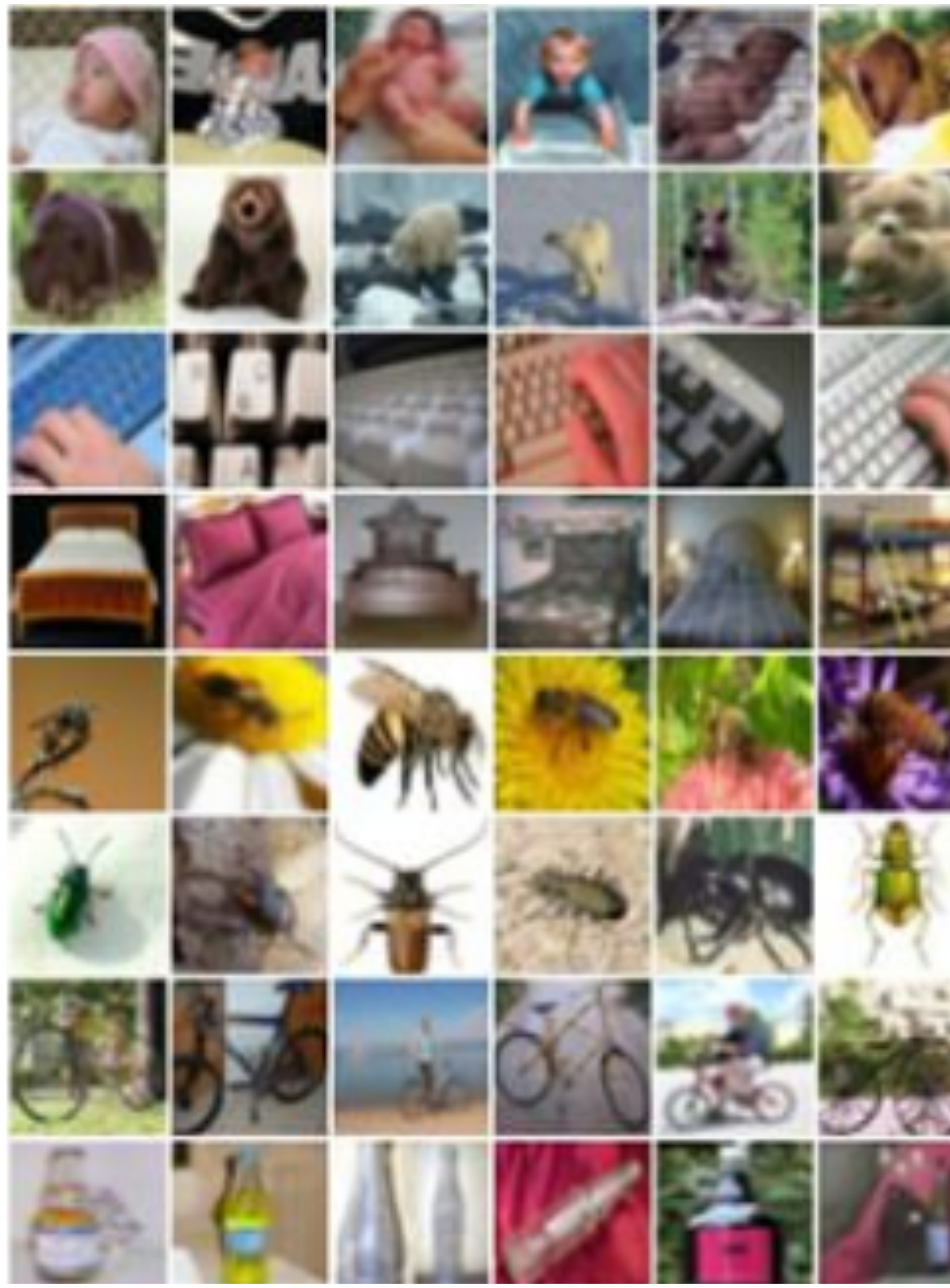
Scientific disciplines & complexity



0 / 1







experiment

scientific datasets
surveys

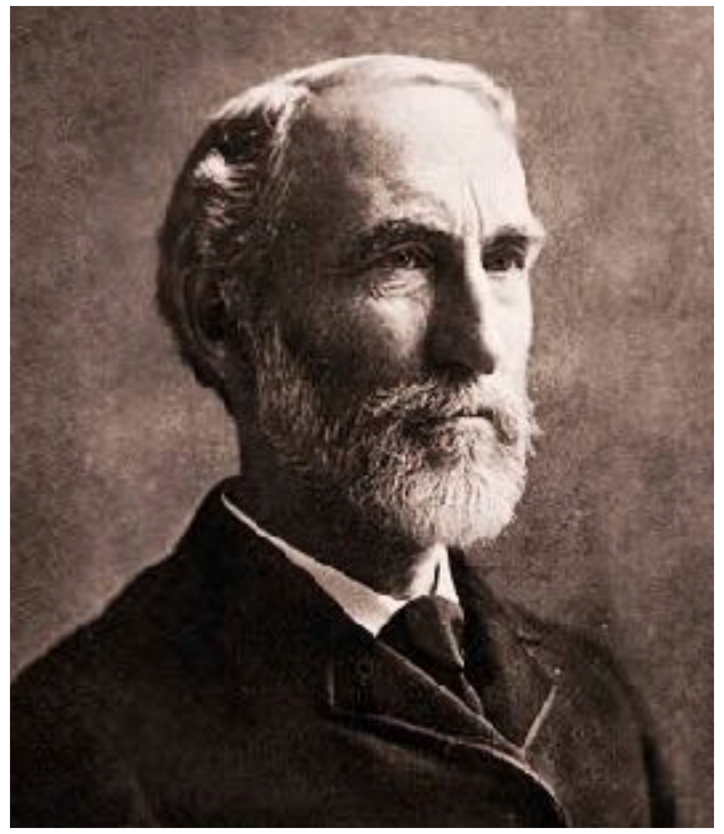
open collection

high

level of controlled variation

low


Model testing / parameter inference
0 / 1
Exploratory data analysis
calibration, control of systematics
experiment
scientific datasets
surveys
open collection
high
level of controlled variation
low



"One of the principal objects of theoretical research is to find
the point of view from which the subject appears in the greatest simplicity."

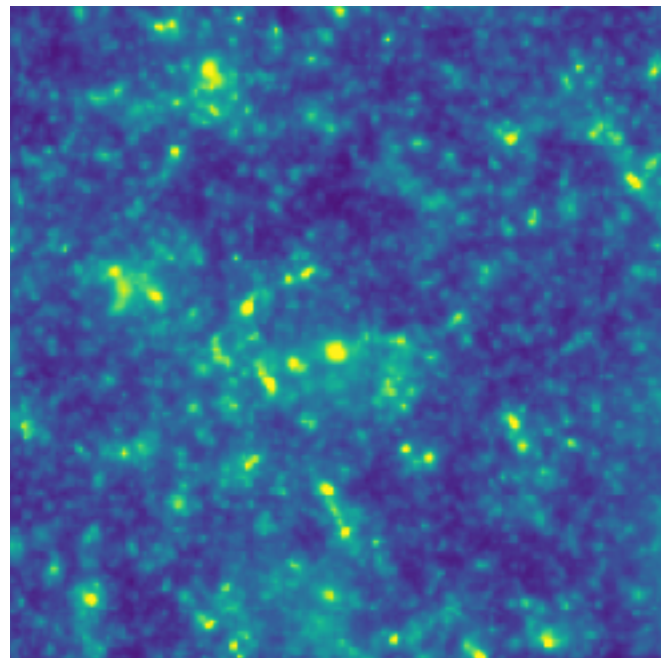
(Gibbs, 1881)

If a system corresponds to a ensemble of states $x = (x_1, \ldots, x_N)$, we want to characterize $P(x)$

For some constraints $\langle \theta_j(x) \rangle = 0$, the Maximum Entropy Principle gives

$$\log P_\theta(x) = \sum_j \lambda_j \theta_j(x) + \text{cst}$$

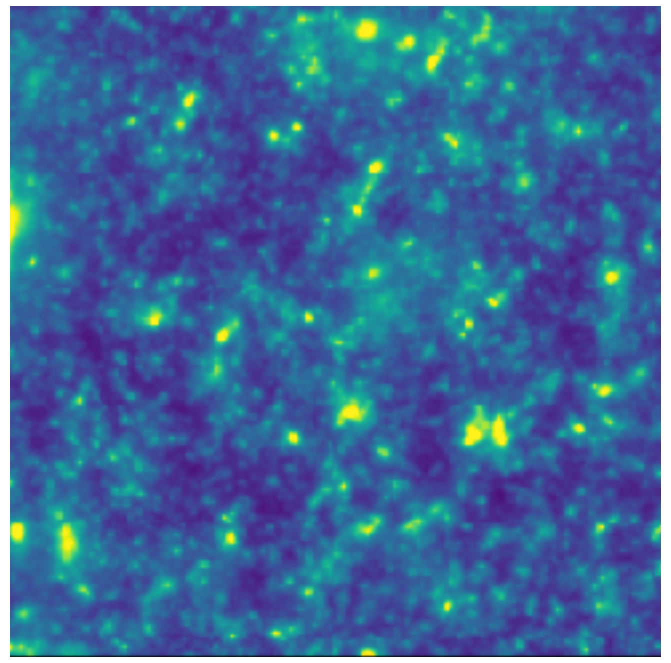
Interacting with data



analysis

statistical description

synthesis

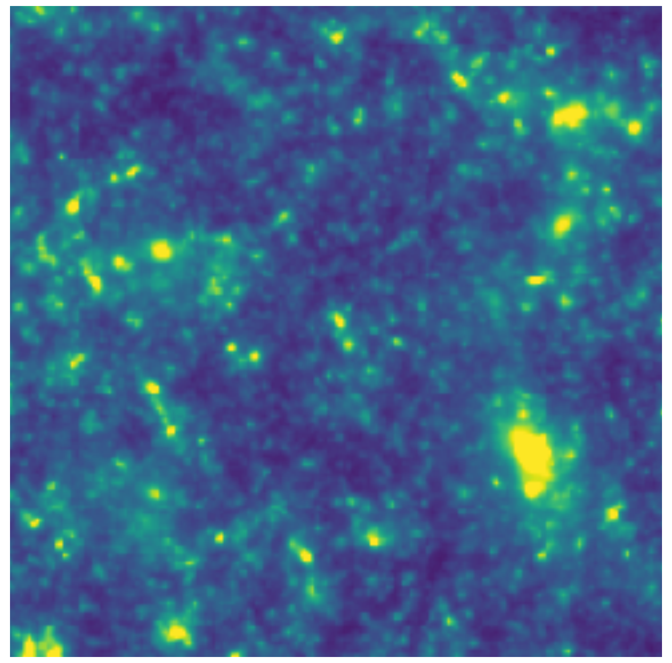


- keep the informative variation
- discard the irrelevant one
- stable
- compact

- communicable
- interpretable

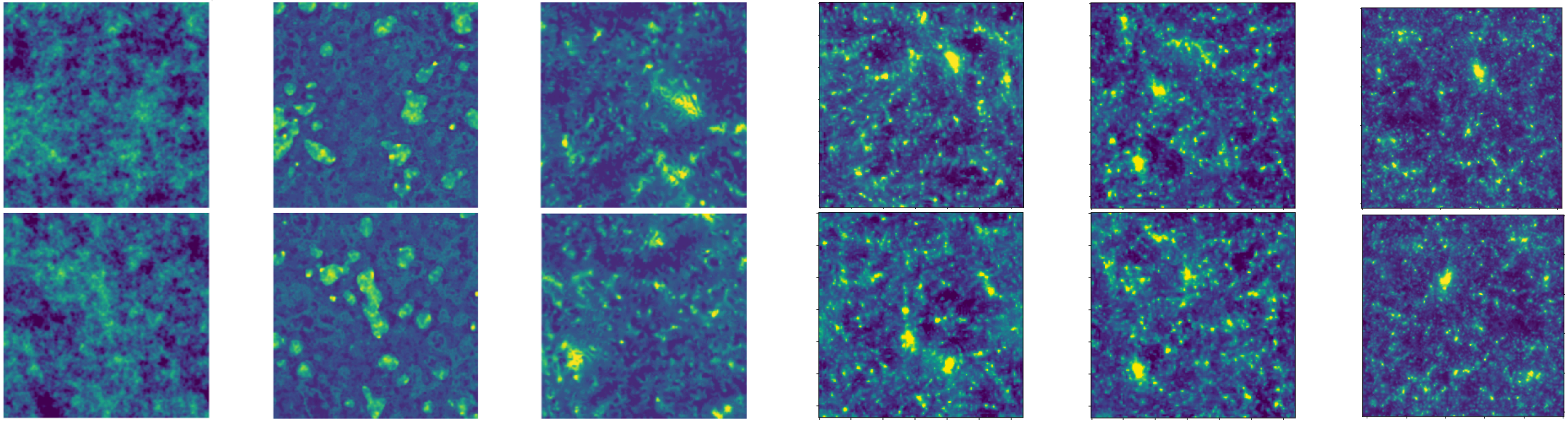
(key properties in science)

Synthesis — generation of data with similar texture

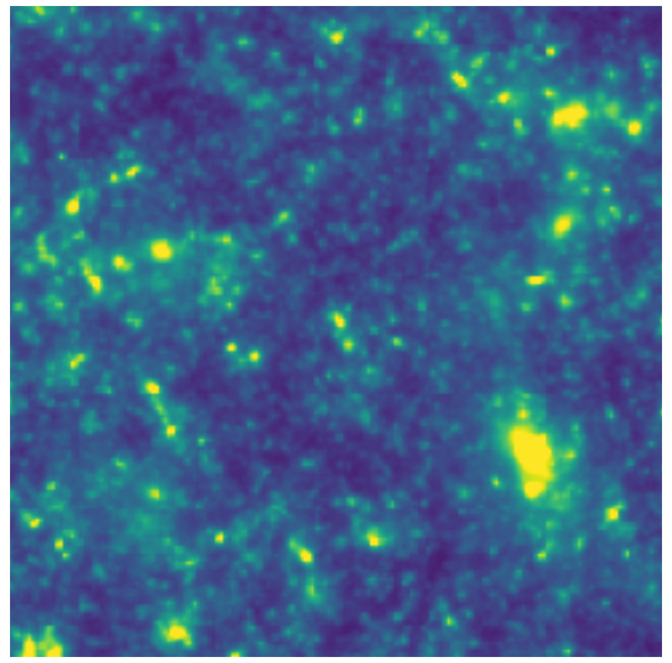


10 | 100 | 1,000 | 10,000 | 100,000 | 1 million

externally trained

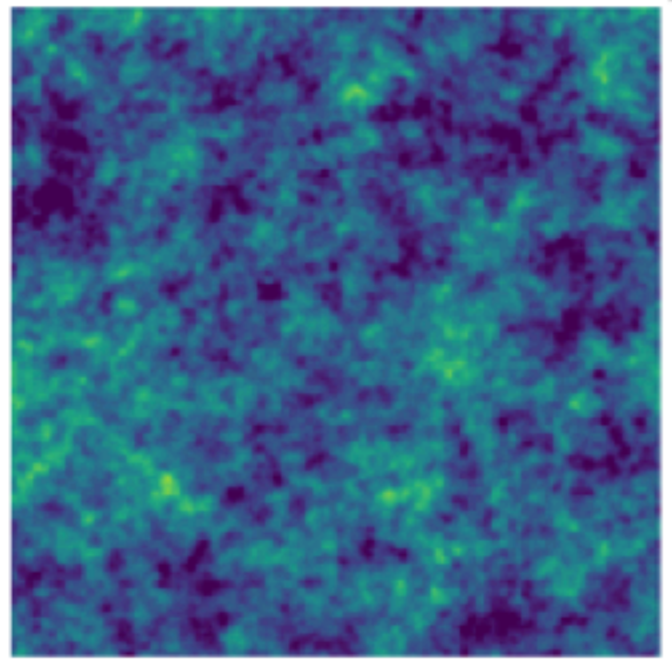


Synthesis — generation of data with similar texture

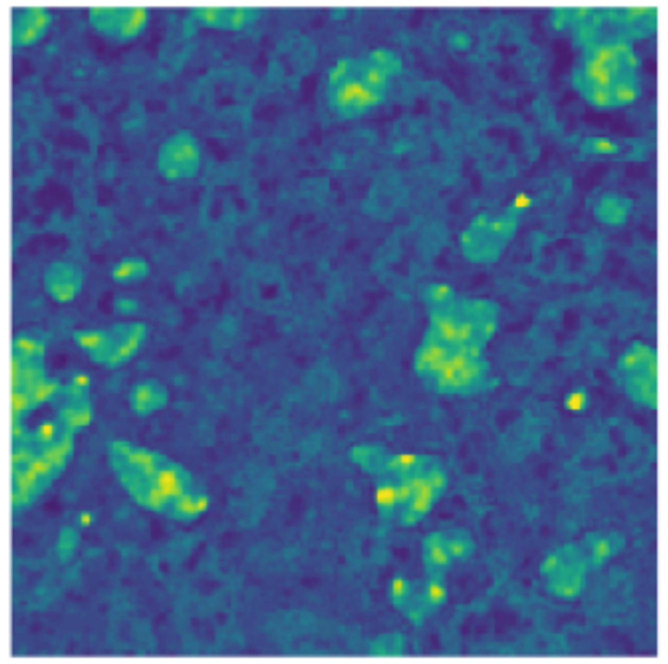


10 | 100 | 1,000 | 10,000 | 100,000 | 1 million

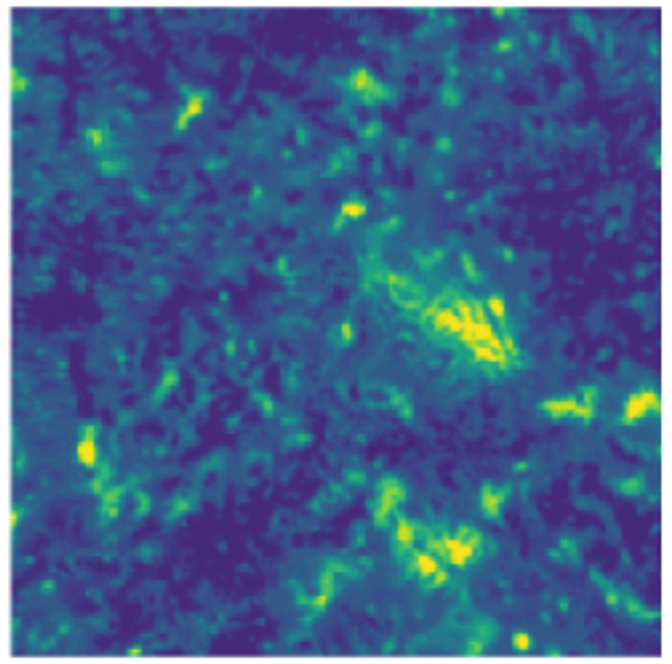
externally trained



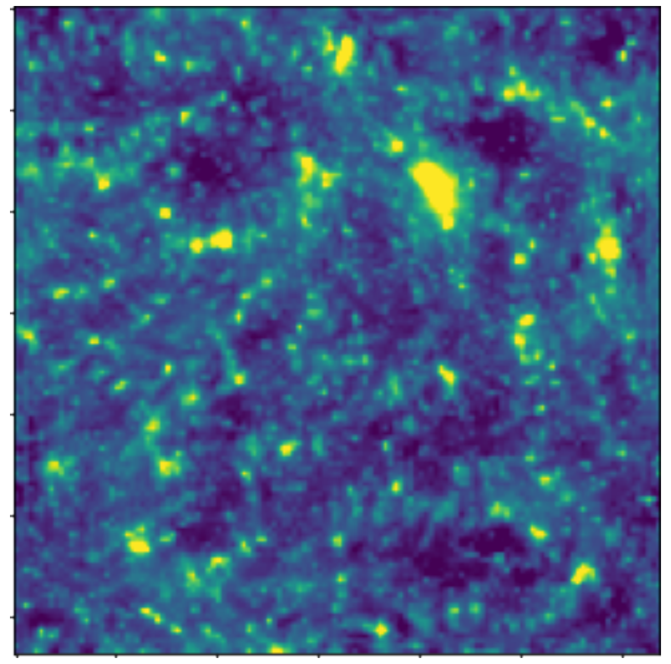
power spectrum



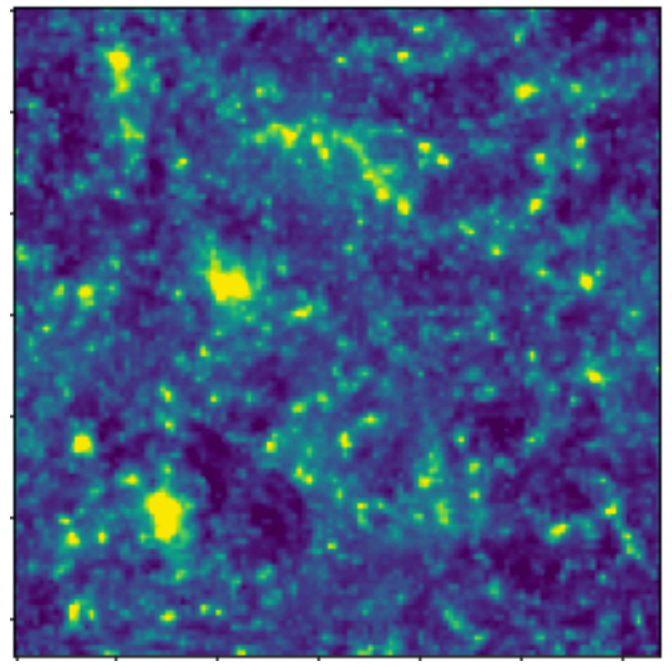
bispectrum



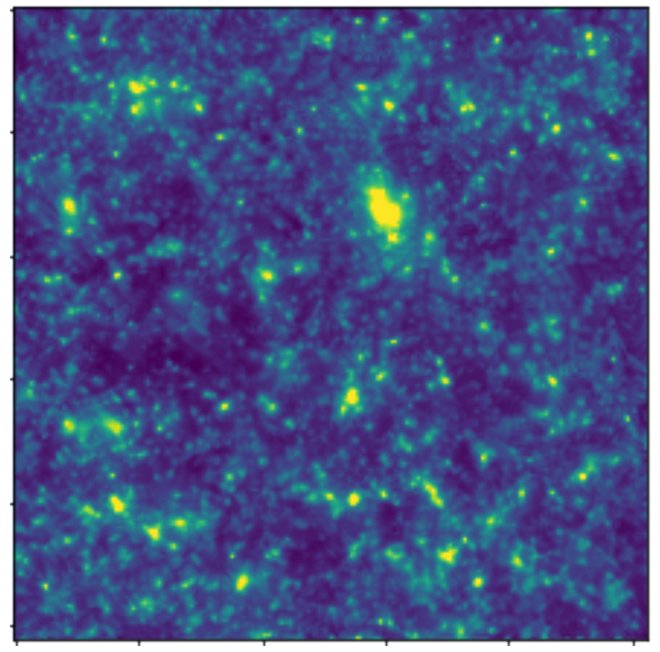
scattering transform



scattering covariance



phase harmonic transform



Convolutional neural network

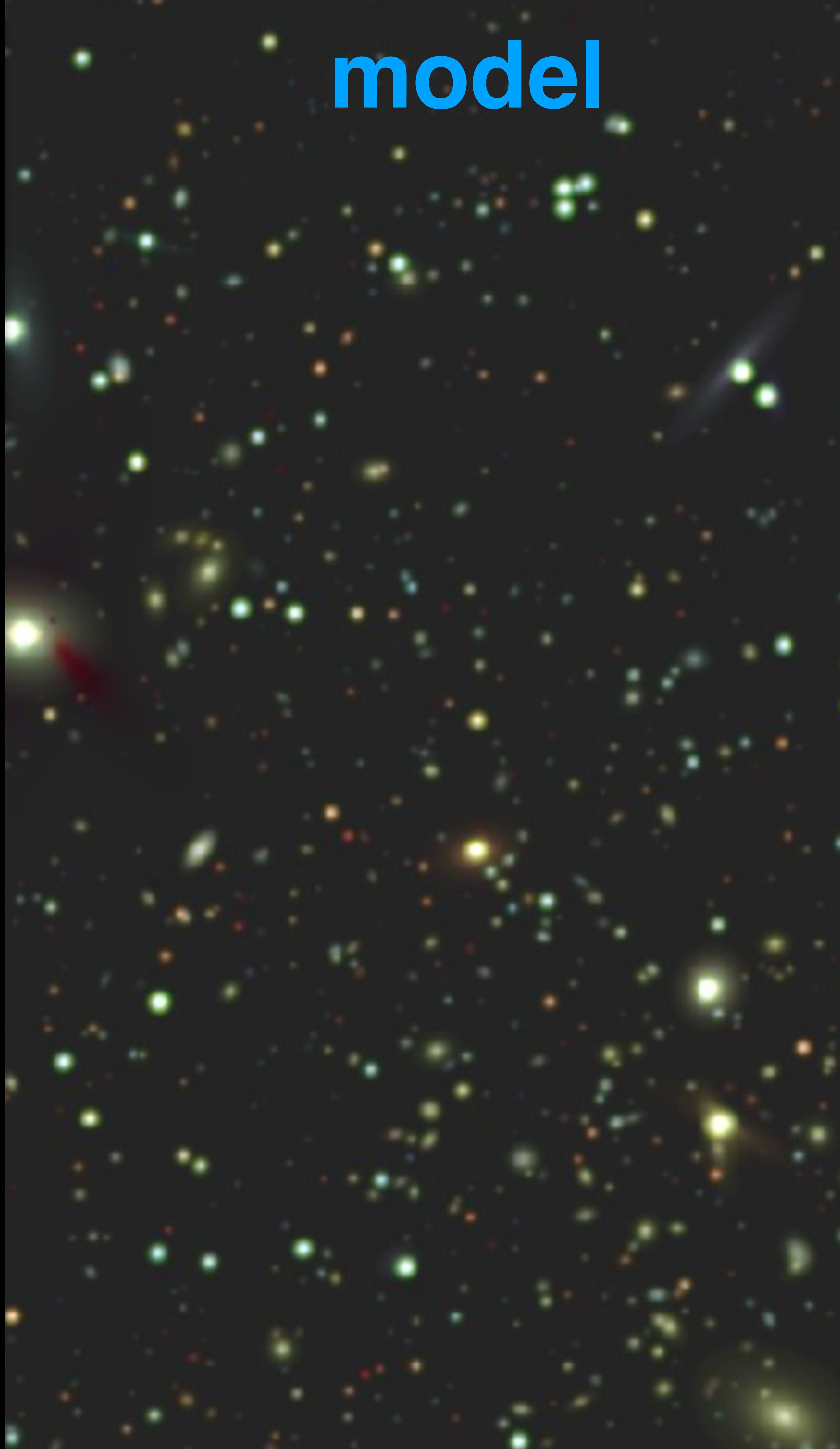

data
model
residuals
RA,Dec = 244.1404, 6.9982, zoom 13
DECaLS DR1 images
DECaLS DR2 images
DECaLS DR2 models
DECaLS DR2 residuals
DECaLS DR1 models
DECaLS DR1 residuals
SDSS images
SFD dust map
Halpha map
unWISE W1/W2
Sources
Bricks
CCDs
Exposures
NGC
Spectra
SDSS Spectro Plates
DECaLS survey
legacysurvey.com

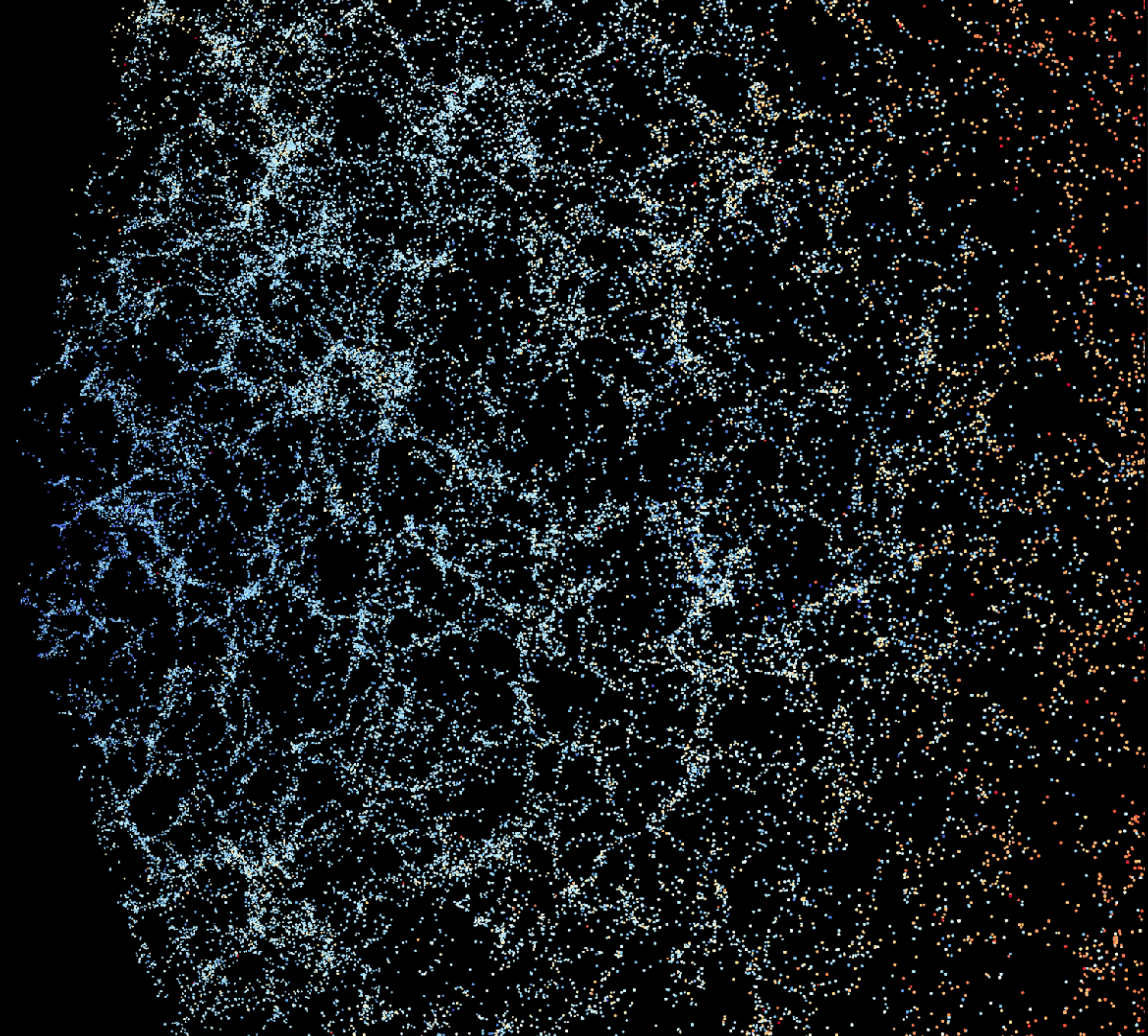

data
model
residuals
RA,Dec = 244.1267, 7.0246, zoom 13
DECaLS DR1 images
DECaLS DR2 images
DECaLS DR2 models
DECaLS DR2 residuals
DECaLS DR1 models
DECaLS DR1 residuals
SDSS images
SFD dust map
Halpha map
unWISE W1/W2
Sources
Bricks
CCDs
Exposures
NGC
Spectra
SDSS Spectro Plates
DECaLS survey
legacysurvey.com

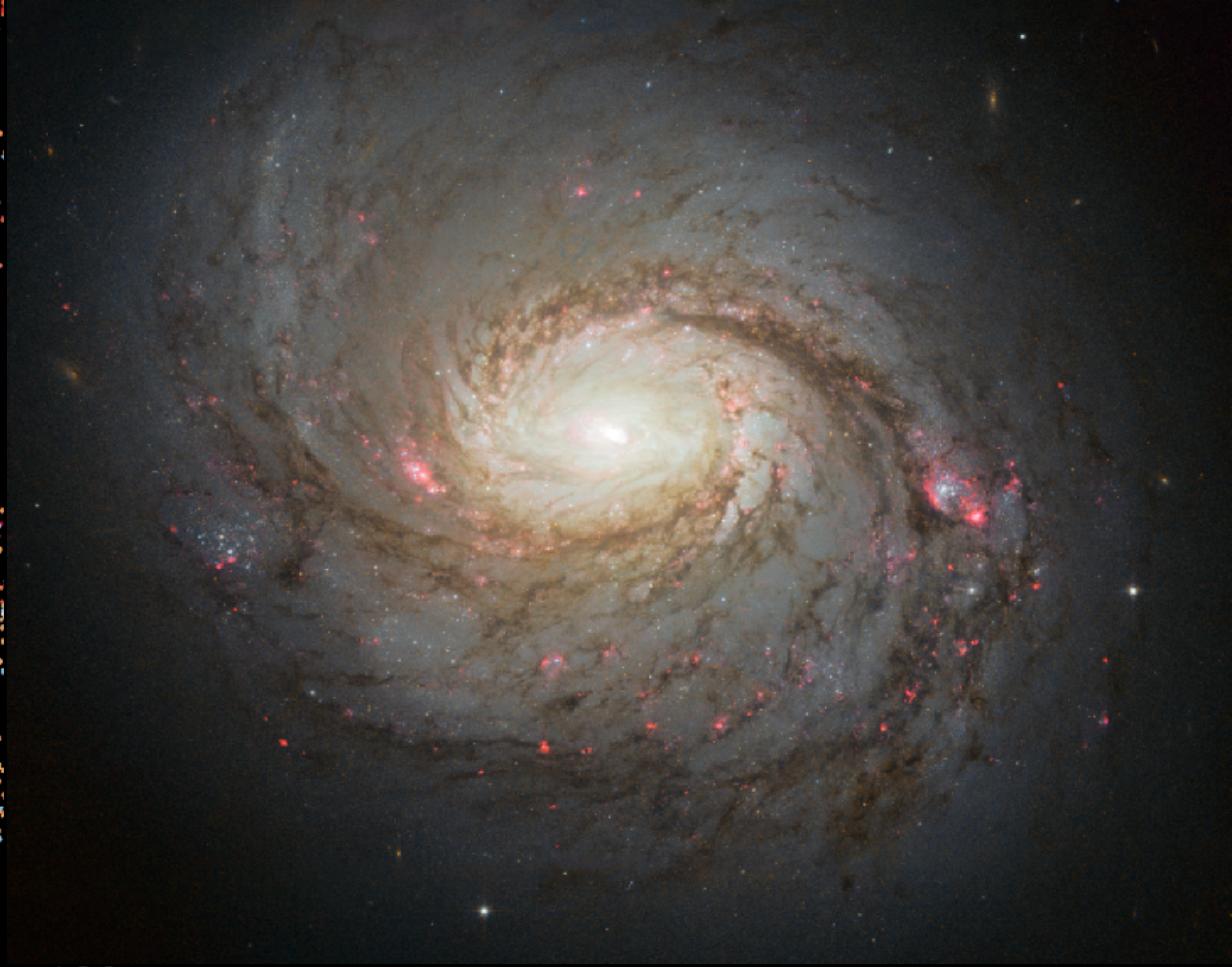

data
model
residuals
RA,Dec = 244.1433, 7.0874, zoom 13
DECaLS DR1 images
DECaLS DR2 images
DECaLS DR2 models
DECaLS DR2 residuals
DECaLS DR1 models
DECaLS DR1 residuals
SDSS images
SFD dust map
Halpha map
unWISE W1/W2
Sources
Bricks
CCDs
Exposures
NGC
Spectra
SDSS Spectro Plates

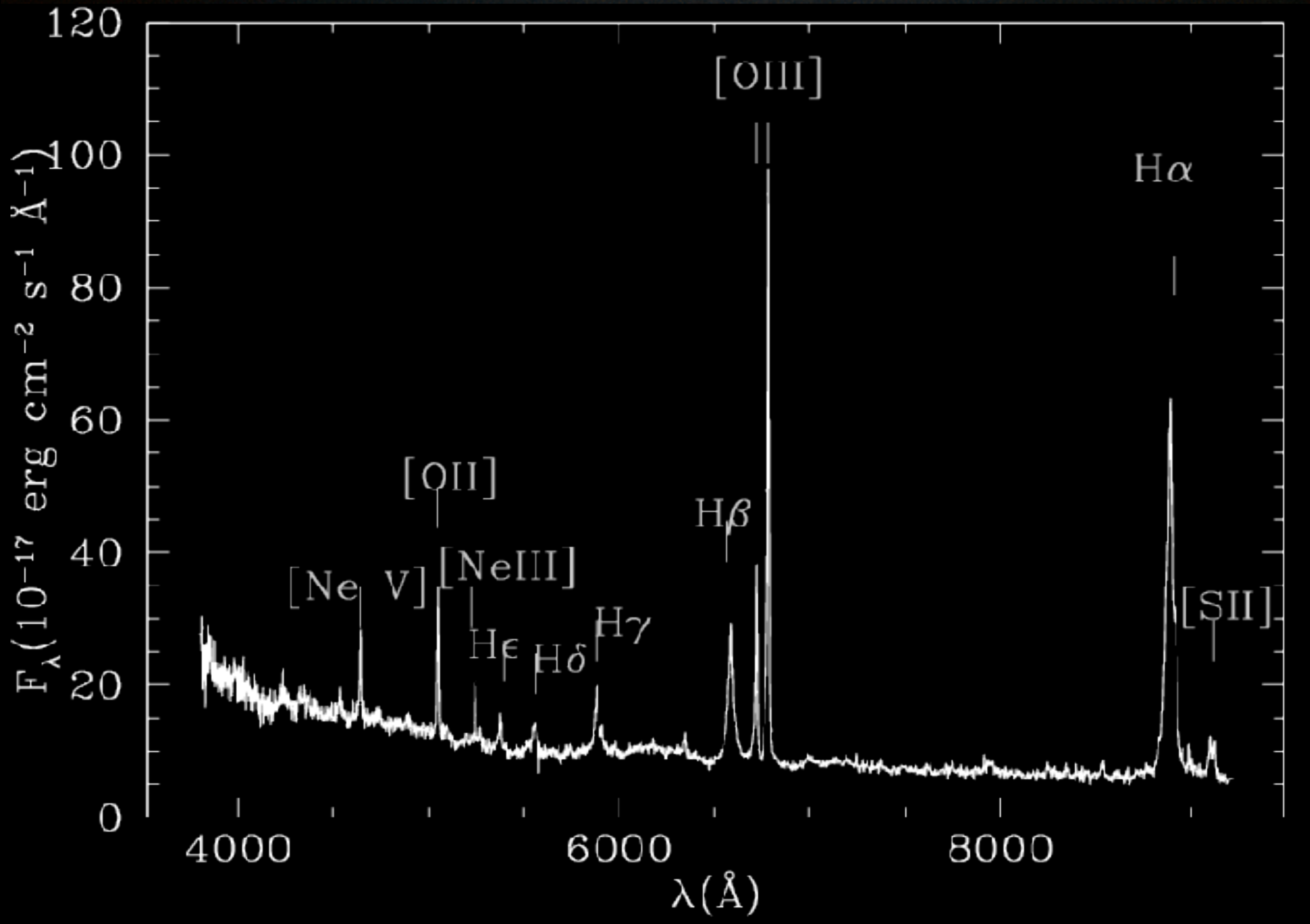
catalog of objects

brightness
position
size
ellipticity
colors
…







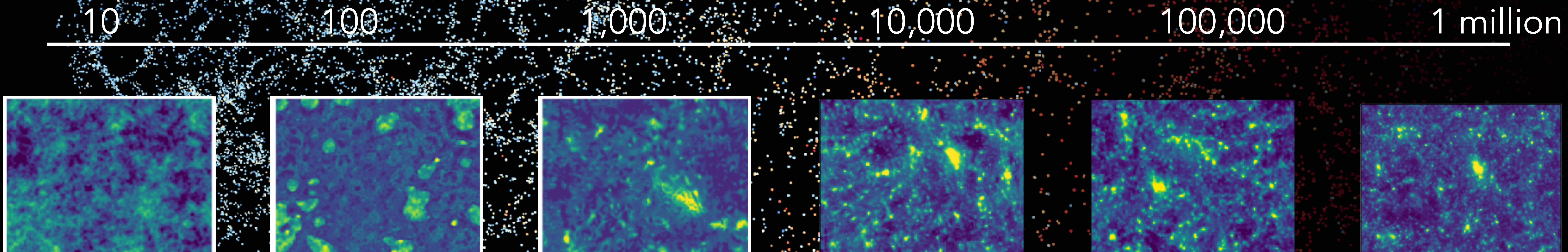

120
100
80
60
40
20
0
$F_\lambda (10^{-17}$ erg cm^{-2} s^{-1} Å$^{-1})$
4000
6000
8000
λ(Å)
[OIII]
Hα
[OII]
[NeIII]
[Ne V]
Hβ
Hϵ
Hδ
Hγ
[SII]

Cosmology

What is the texture of the Universe?

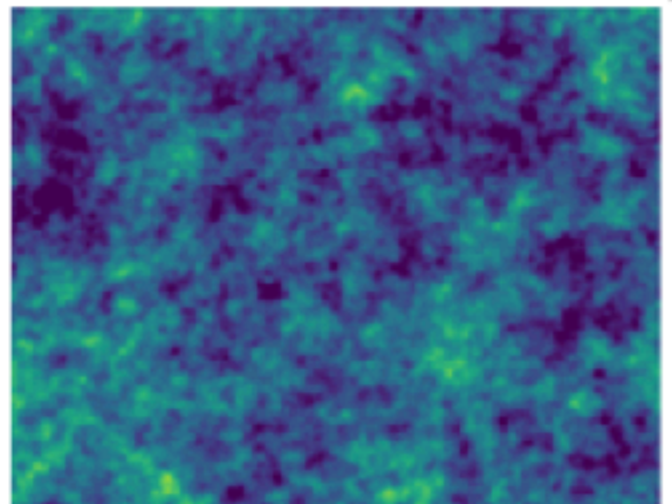
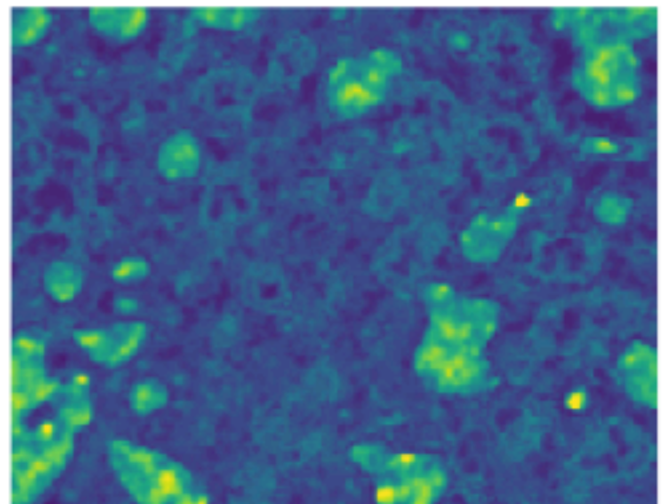
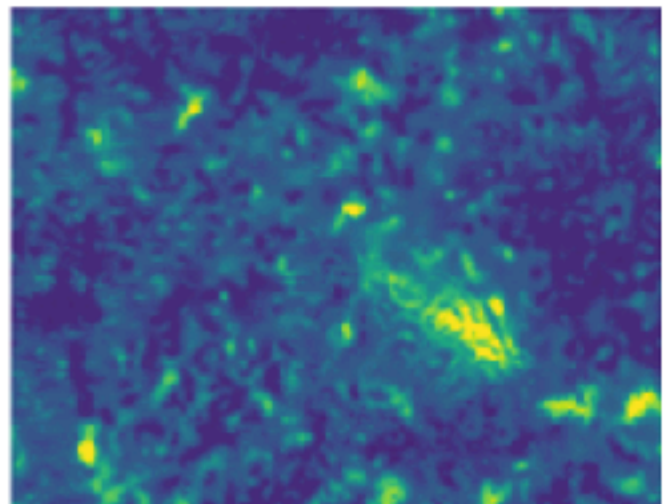
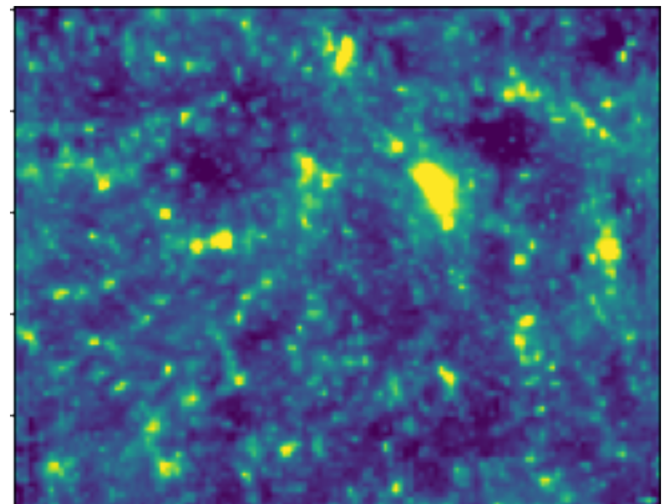
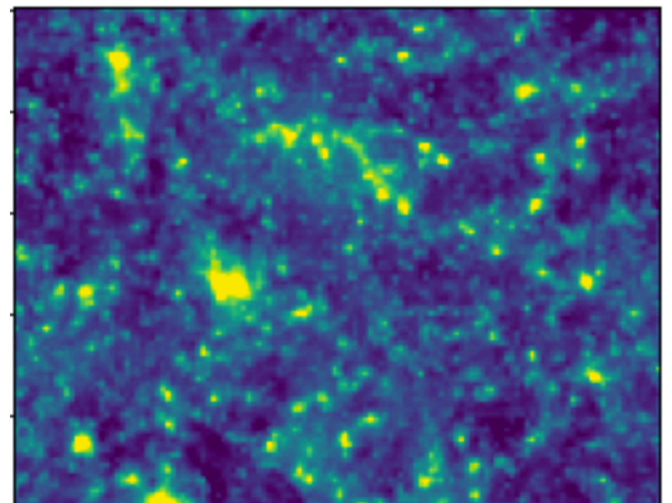
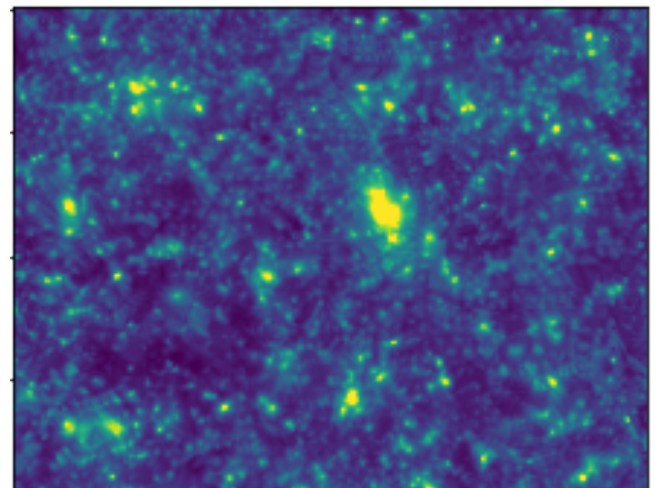
- density?
- amplitude of fluctuations?
- rate of expansion?

→ the cosmological parameters



stationary fields or texture

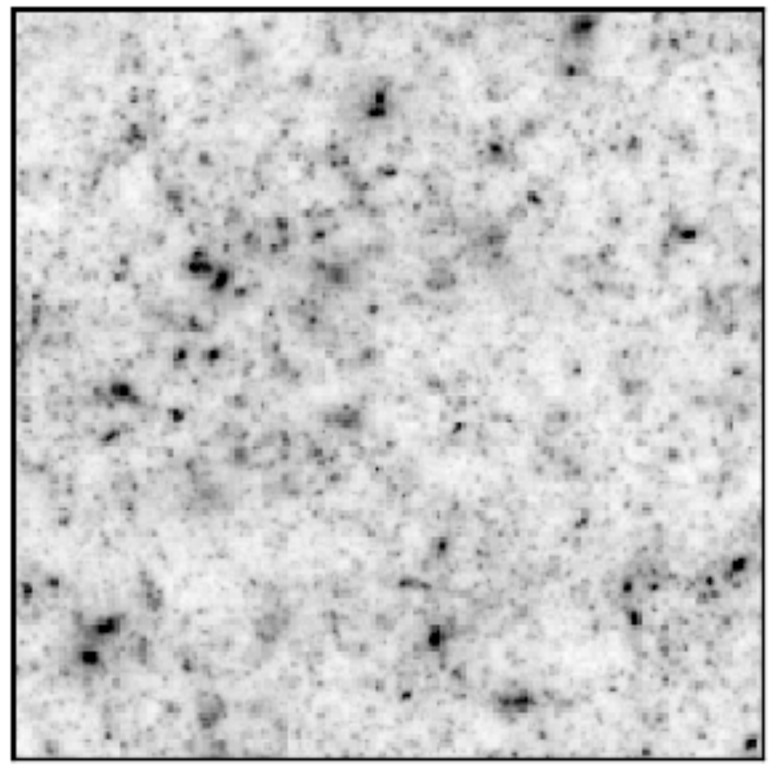
10 100 1,000 10,000 100,000 1 million

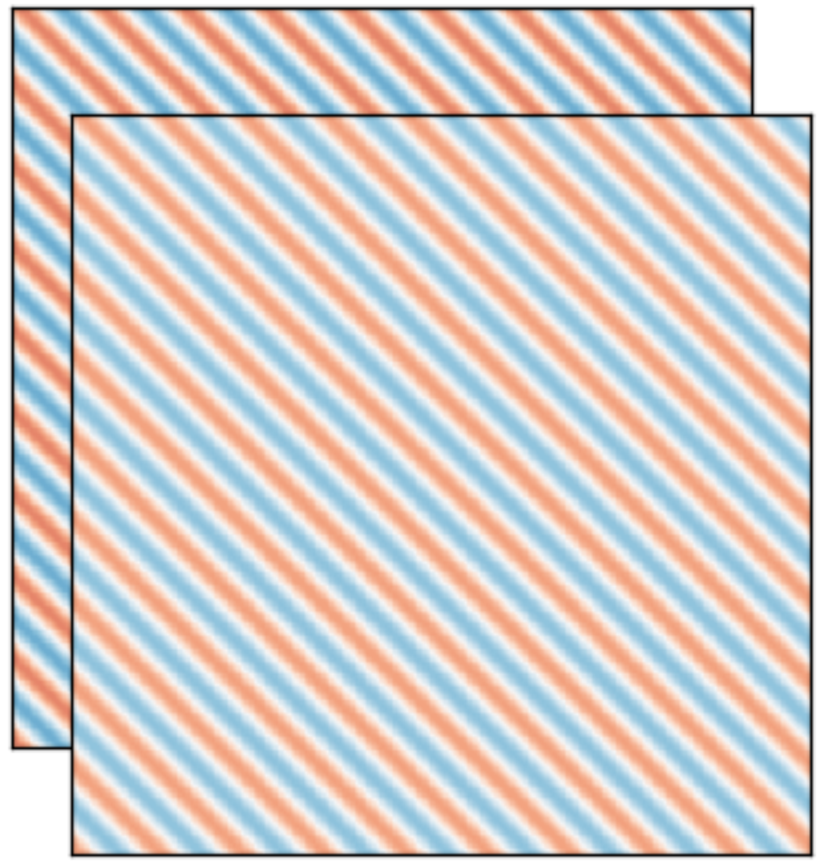
Power spectrum

Let's pick a frequency $\overrightarrow{k}_1$ $\quad P(k_1) = \langle |I_0 * \mathrm{e}^{i k_1 x}|^2 \rangle$

I_0



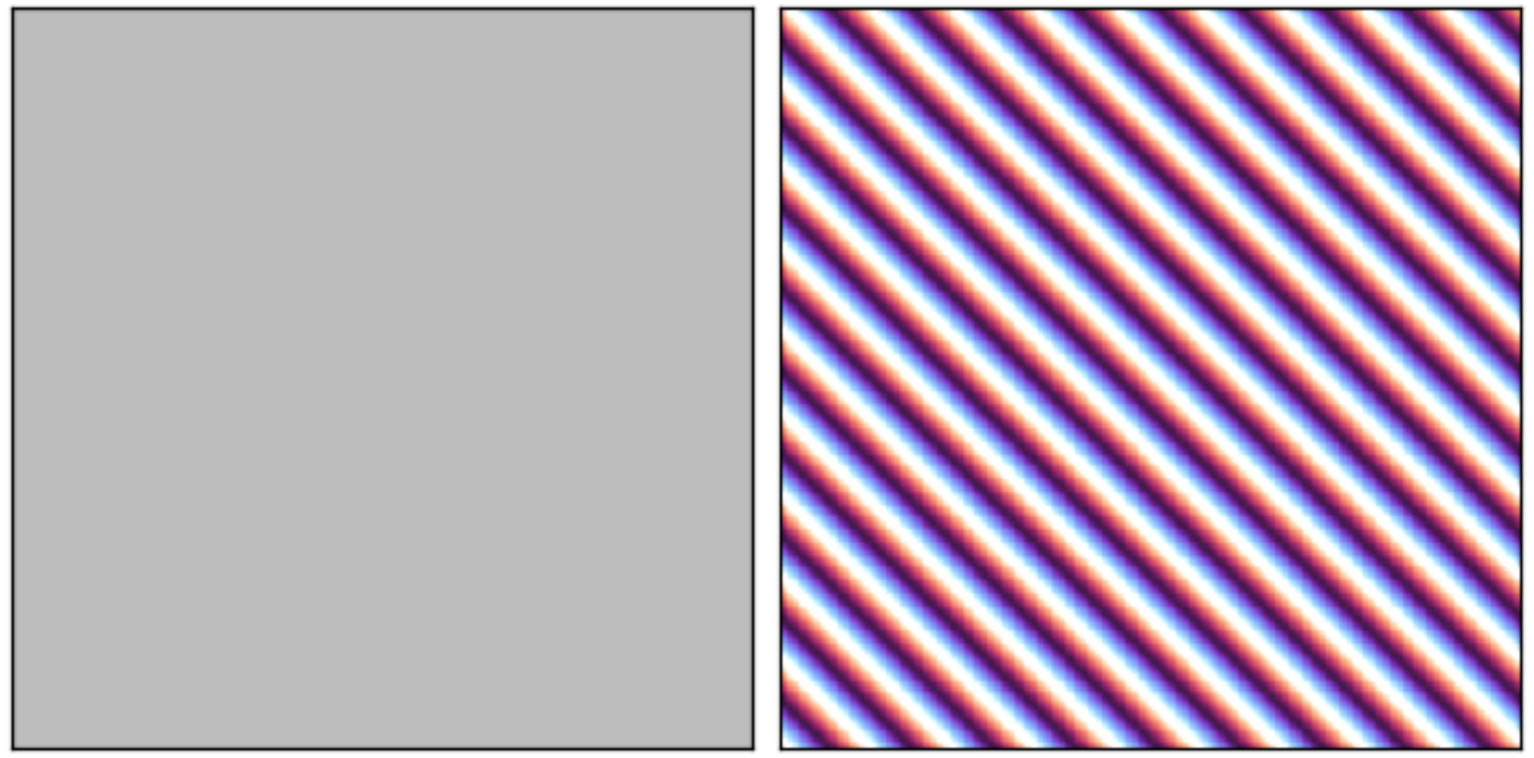
convolution by $e^{i k_1 x}$



real & imaginary

=

discarded



modulus phase

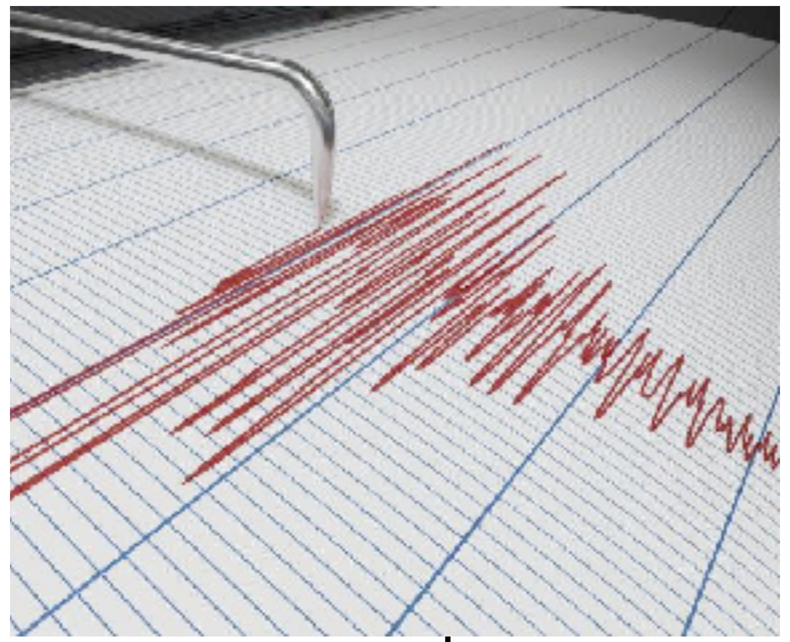
L_2 norm $= P(k_1)$

$P(k)$ **properties:**

- Invariant to translation (& possibly rotations)
- Conservation of energy
- Separation of scales
- All information extracted if Gaussian random field

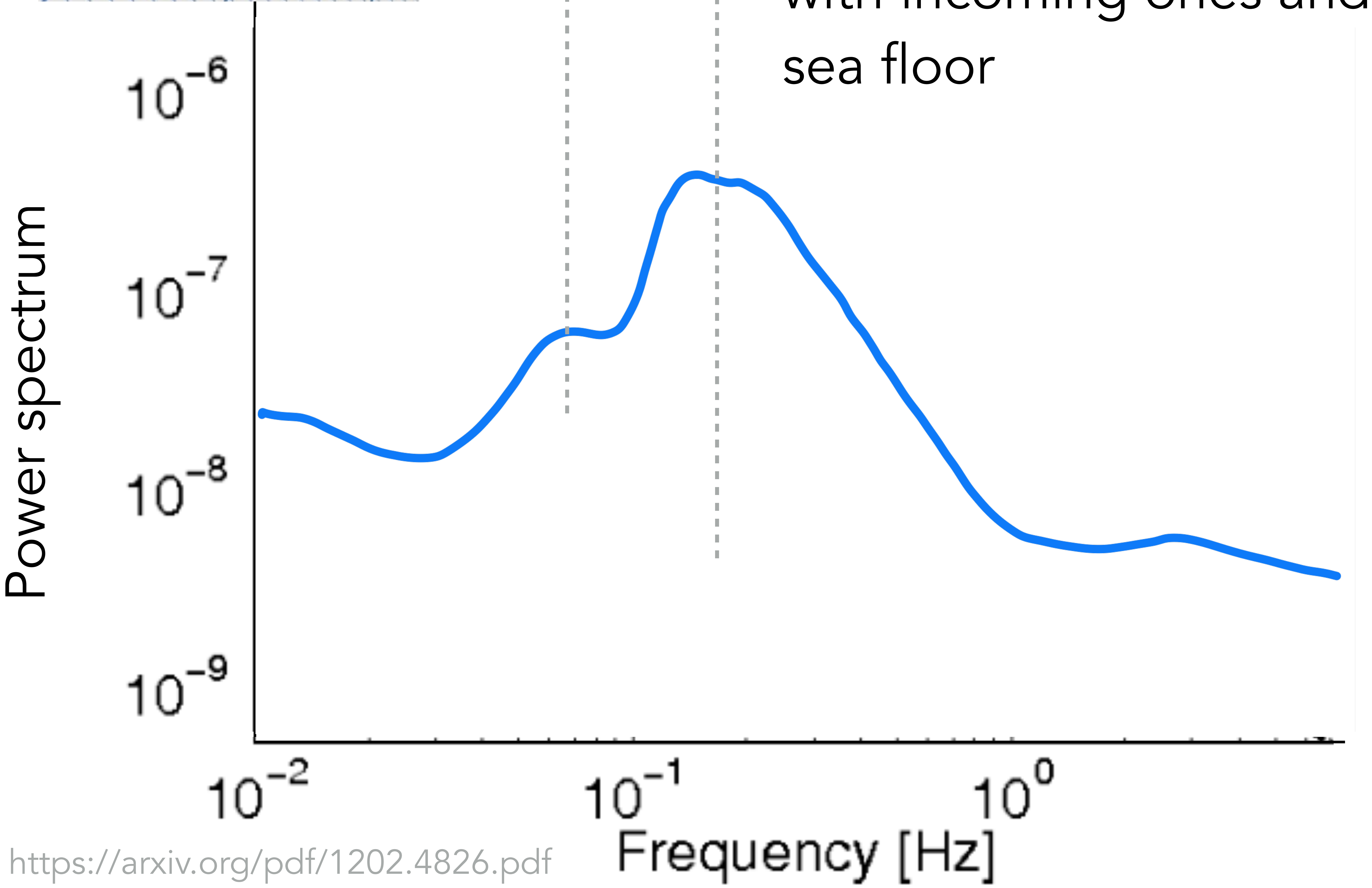
1D power spectrum

seismogram

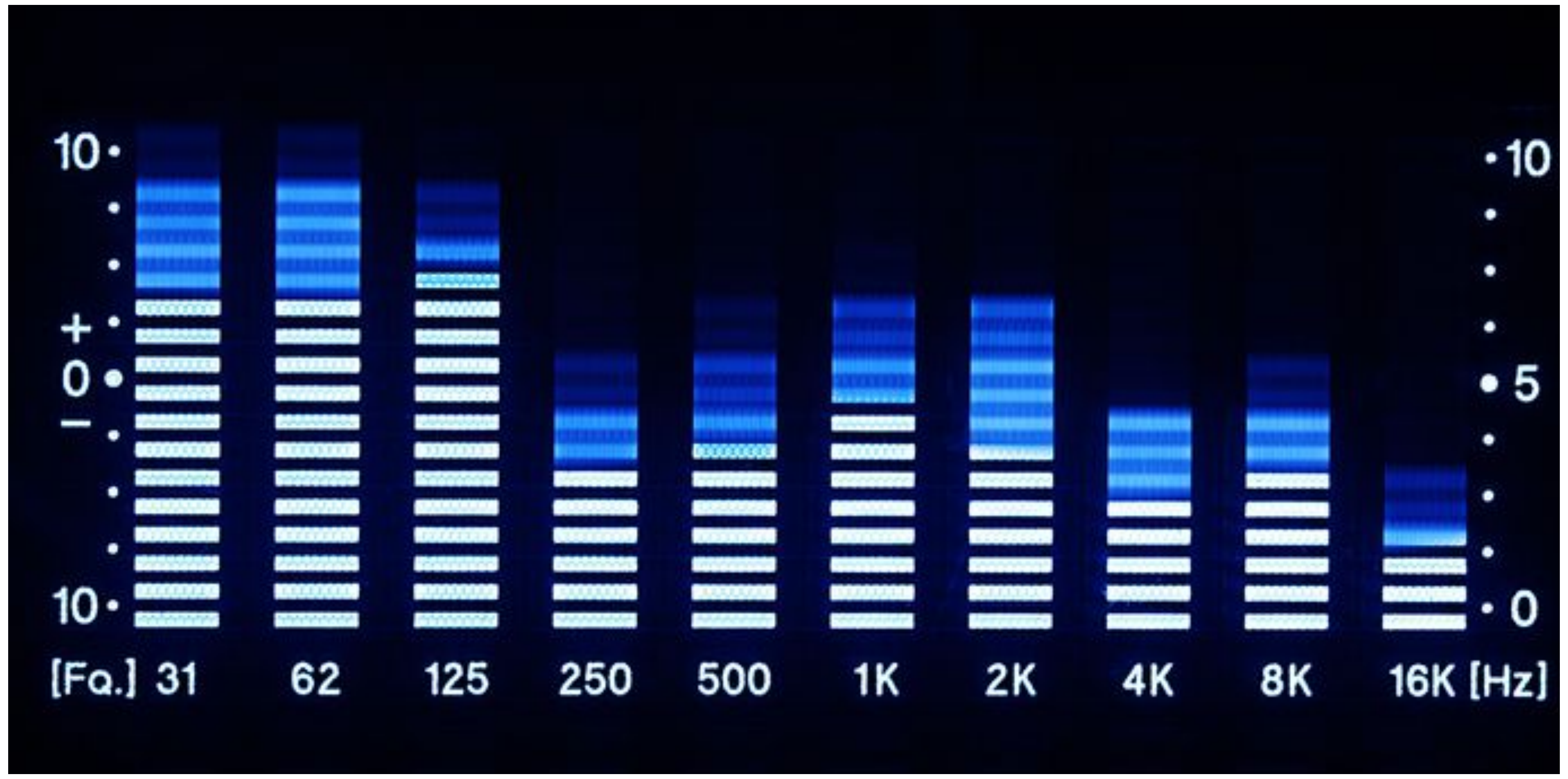


~15s: ocean waves hitting coastlines

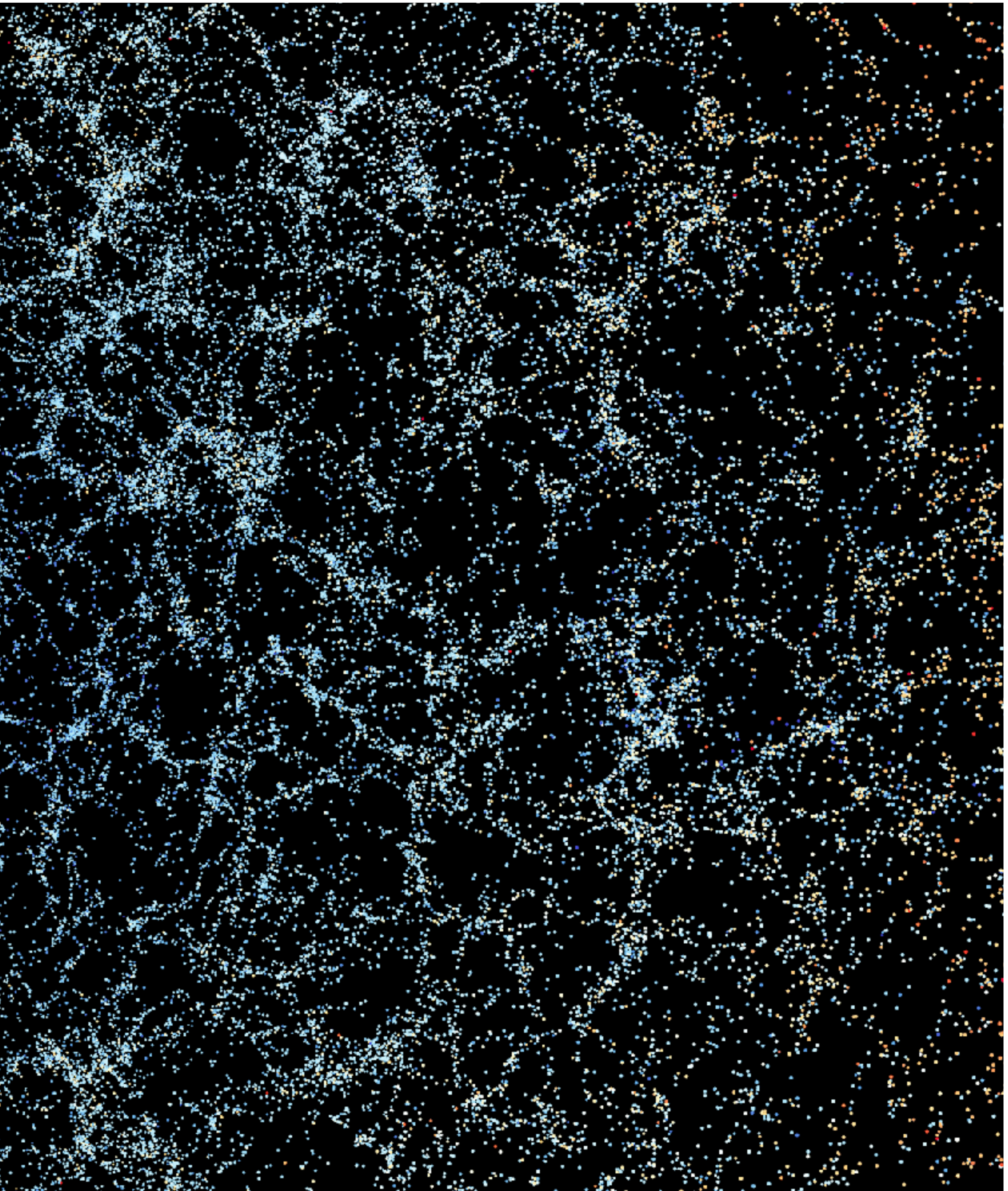
~7s: reflected waves interacting with incoming ones and the sea floor

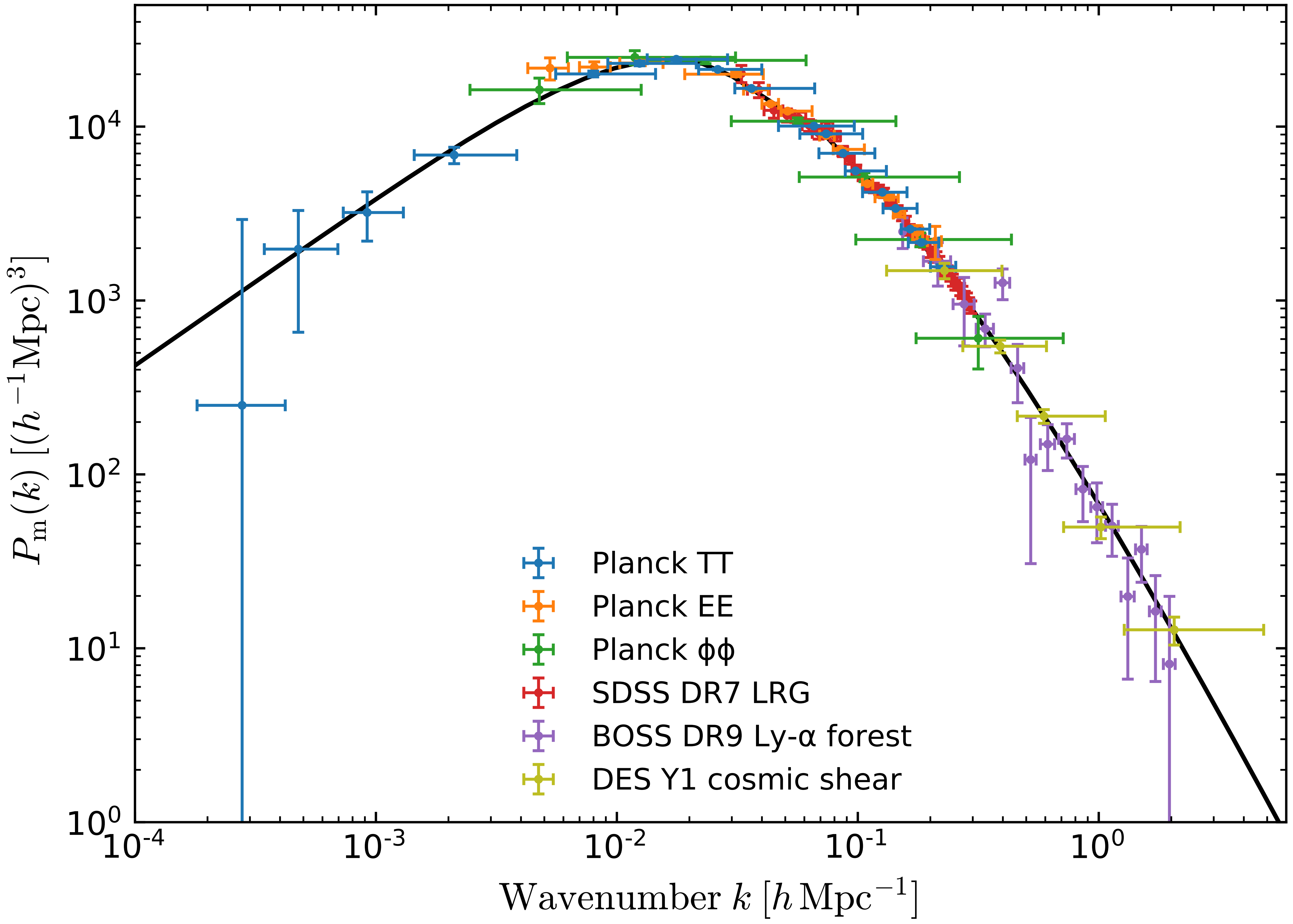


https://arxiv.org/pdf/1202.4826.pdf



3D power spectrum in cosmology

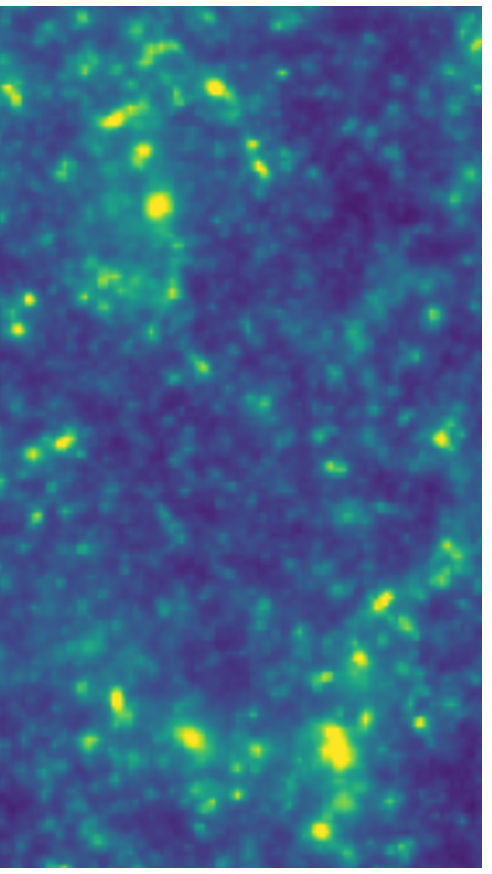




> the gravitational potential energy (per unit volume)

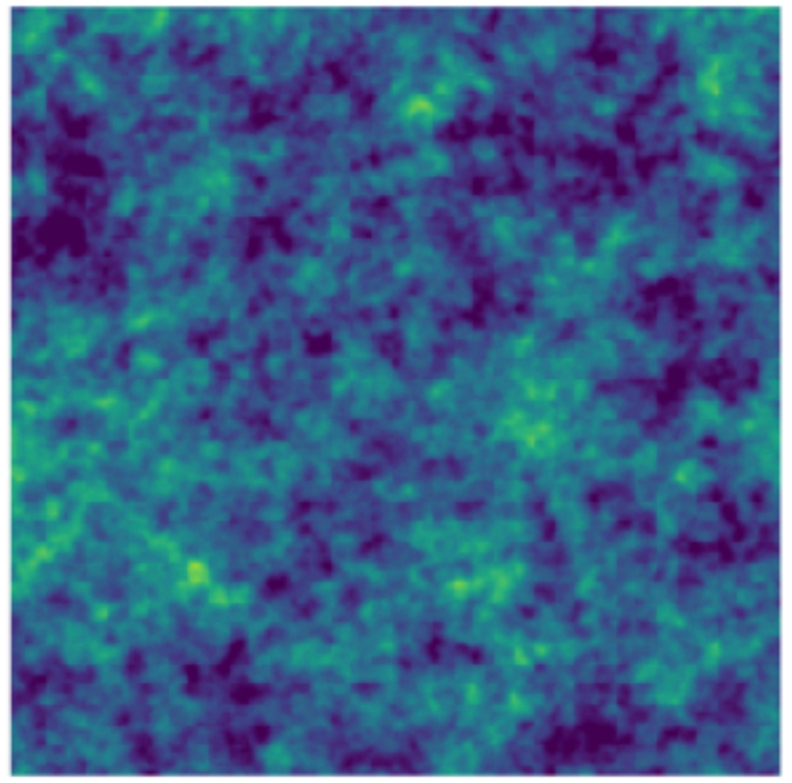
$$W = -\frac{3\Omega_{\rm m} H_0^2}{8\pi^2 a}\int_0^\infty {\rm d}k\ P(k,a)$$

The limitations of moment-based approaches



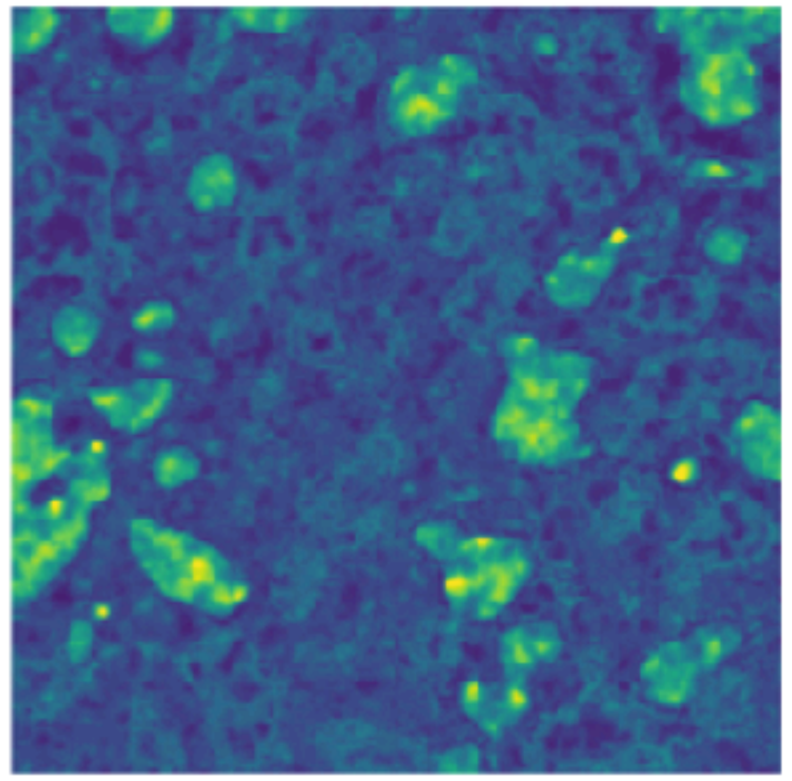
Power spectrum

$$P(k) = \langle \, | I_0 * \mathrm{e}^{ikx} |^2 \rangle$$

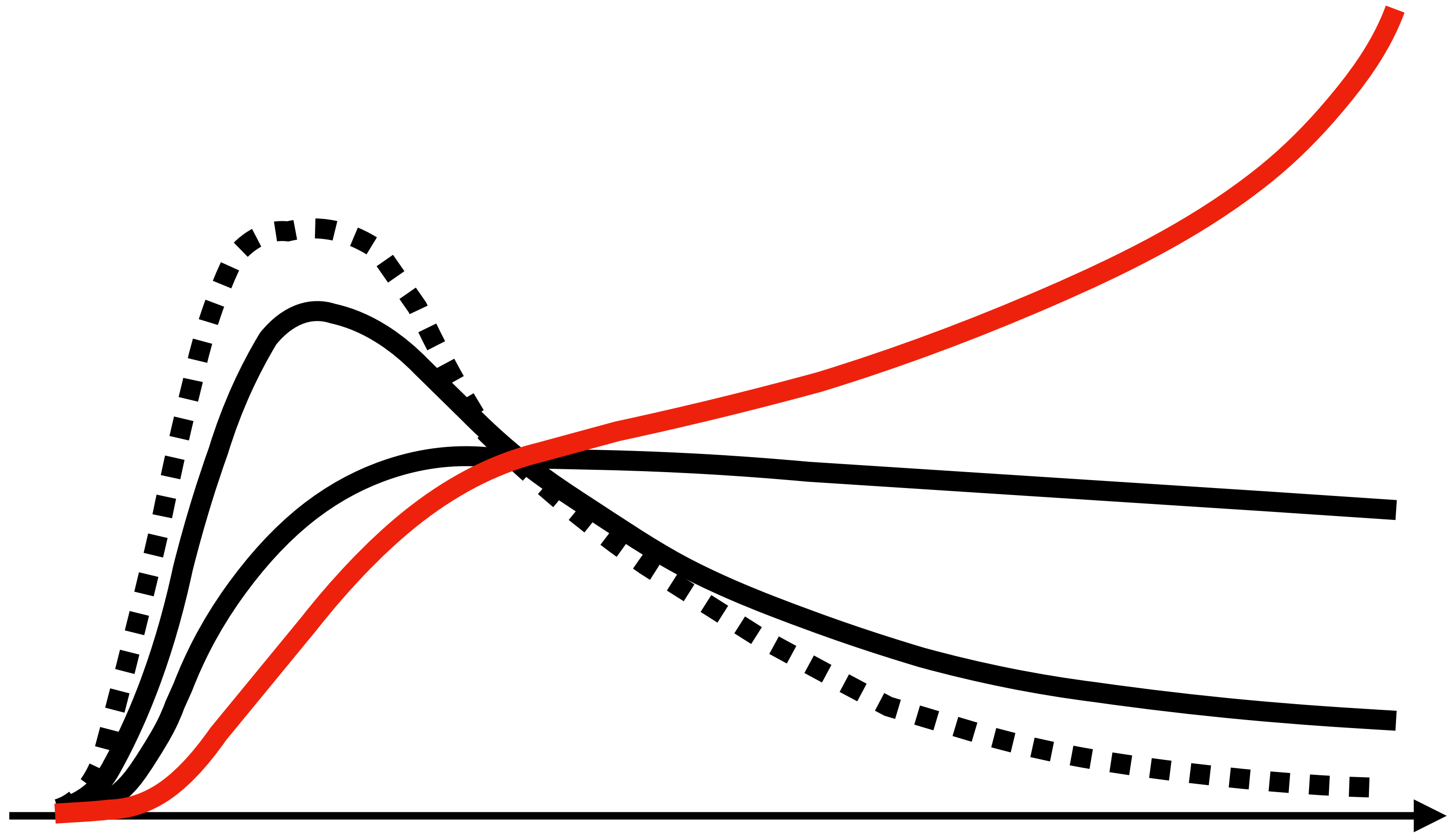


Higher-order statistics

$$B(k_1, k_2, k_3) = \langle (I_0 * \mathrm{e}^{ik_1x}) \; (I_0 * \mathrm{e}^{ik_2x}) \; (I_0 * \mathrm{e}^{ik_3x}) \rangle$$



High-order moments amplify the tail: $\langle \, x \, \rangle, \langle \, x^2 \, \rangle, \langle \, x^3 \, \rangle, \ldots$

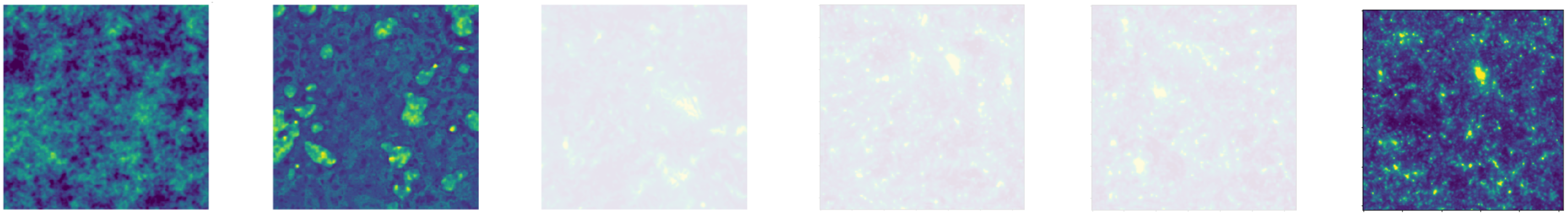


motivated by physics,
perturbation theory
but too limited and unstable

?

driven by performance
on complex tasks
but "black boxes"

10 | 100 | 1,000 | 10,000 | 100,000 | 1 million



power spectrum

bispectrum

scattering transform

scattering covariance

phase harmonic transform

Convolutional neural network

motivated by physics,
perturbation theory
but too limited and unstable

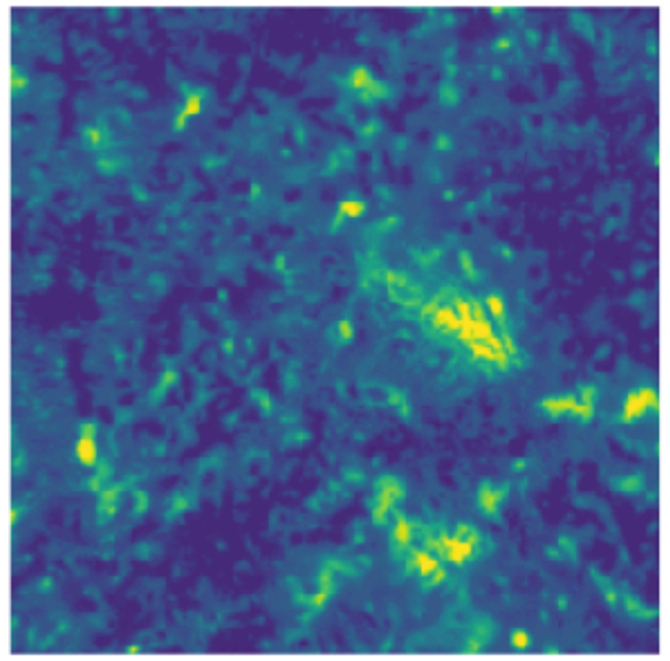
motivated by the mathematics
of neural networks

driven by performance
on complex tasks
but “black boxes”

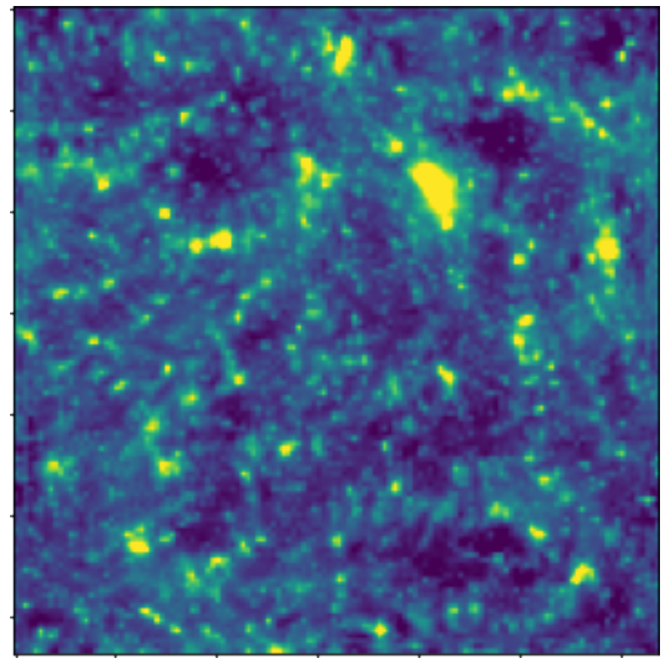
10 | 100 | 1,000 | 10,000 | 100,000 | 1 million

power
spectrum

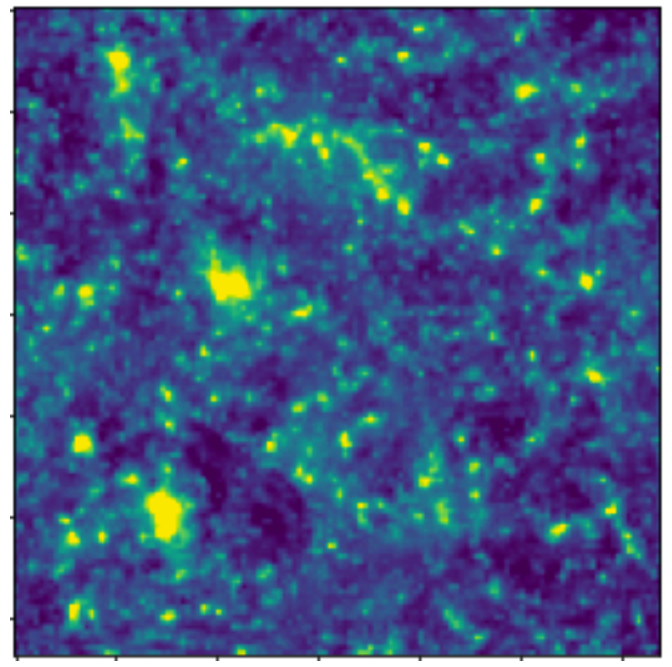
bispectrum



scattering
transform



scattering
covariance

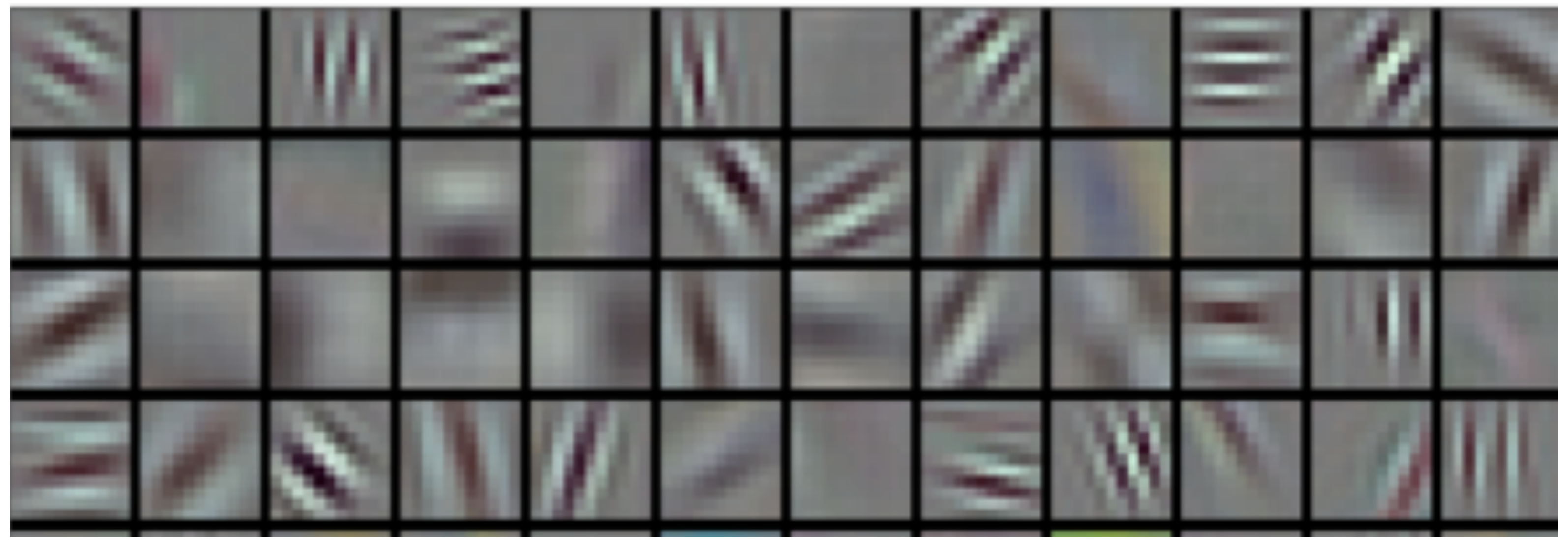


phase harmonic
transform

Convolutional
neural network

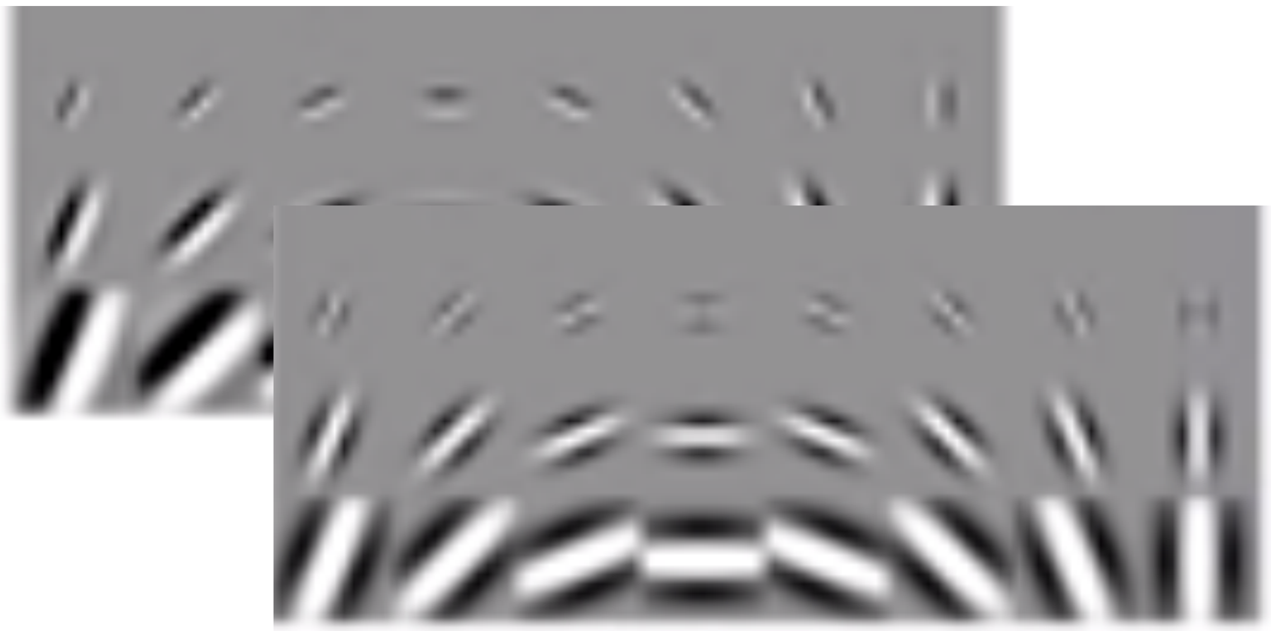
How to design a mathematical network?

kernels learned in AlexNet



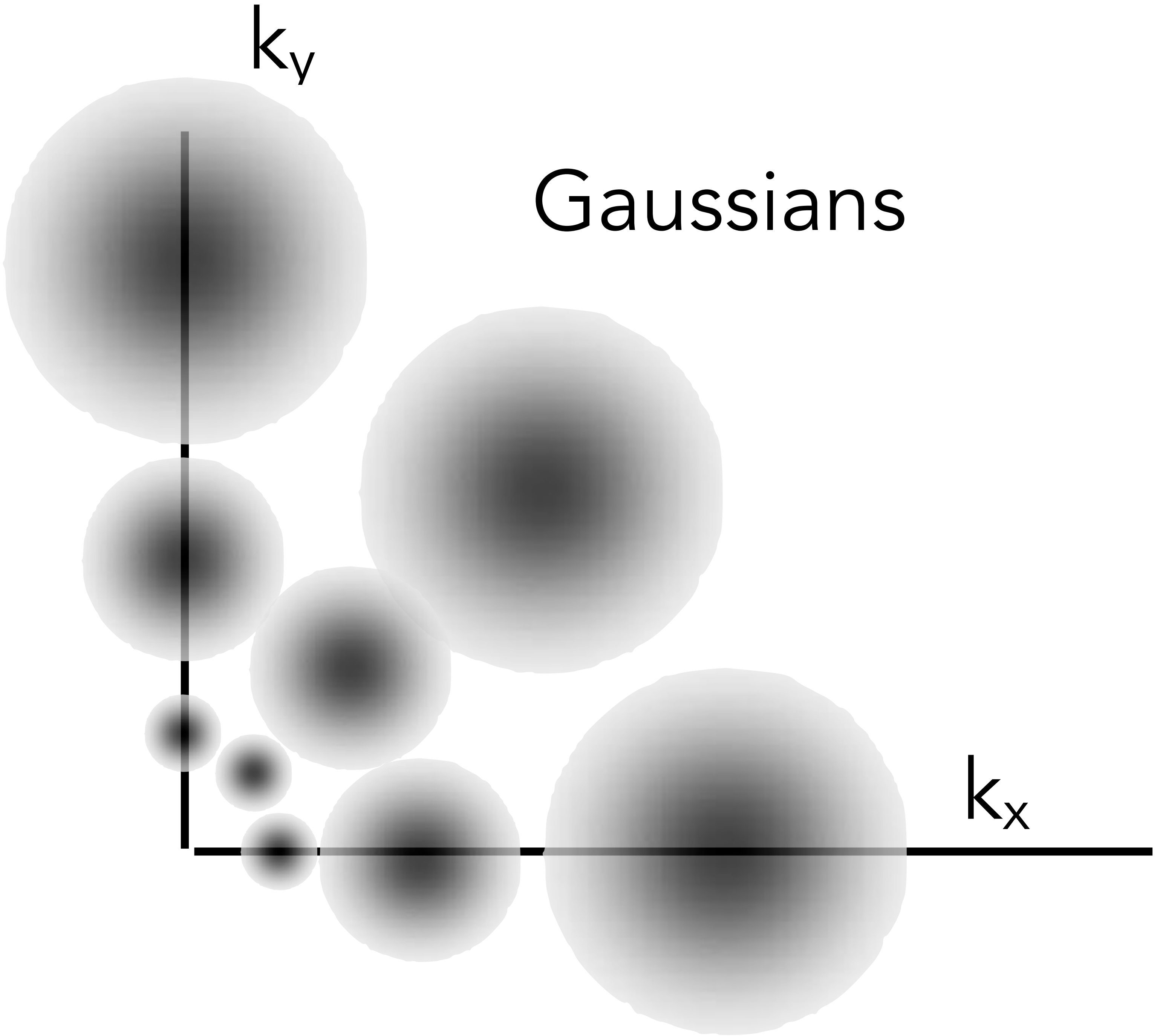
(Krizhevsky, Sutskever, & Hinton 2012)

Let's organize them



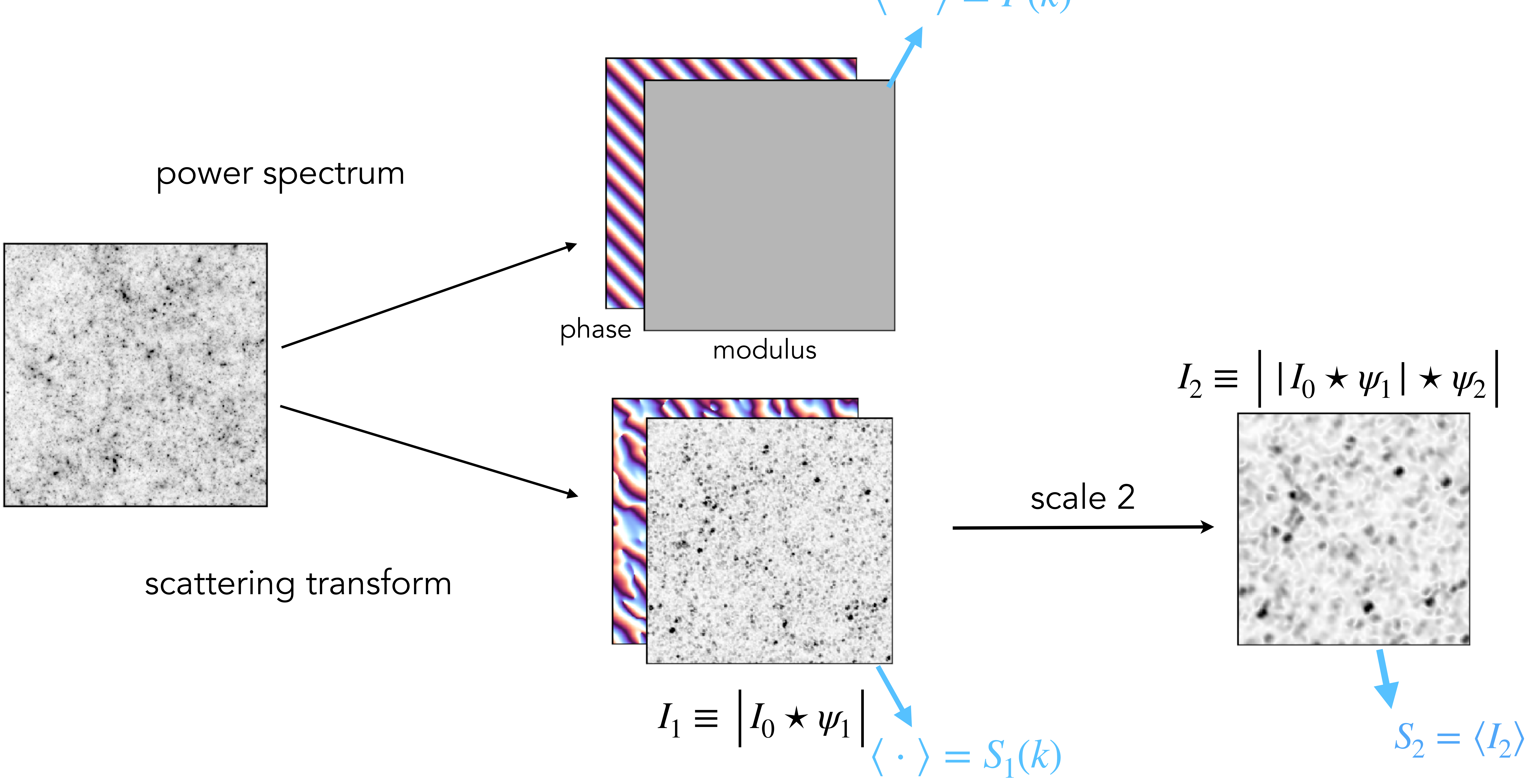
→ Gabor wavelets $\psi(x)$

Fourier representation



→ family of scaled & rotated Gaussians $\tilde{\psi}(k)$

Power spectrum vs scattering transform



Scattering transform

Mallat 2012

$$S_1(k_1) \;= \langle \left| I \star \psi \right| \rangle$$

$$S_2(k_1, k_2) \;= \langle \,|\,|I * \psi_{k_1}| * \psi_{k_2}|\, \rangle$$

$$\dots$$

$$S_n(k_1, \dots, k_n) \;= \langle \,|\,|\,|I * \psi_{k_1}| * \psi_{k_2}| \dots * \psi_{k_n}|\, \rangle$$

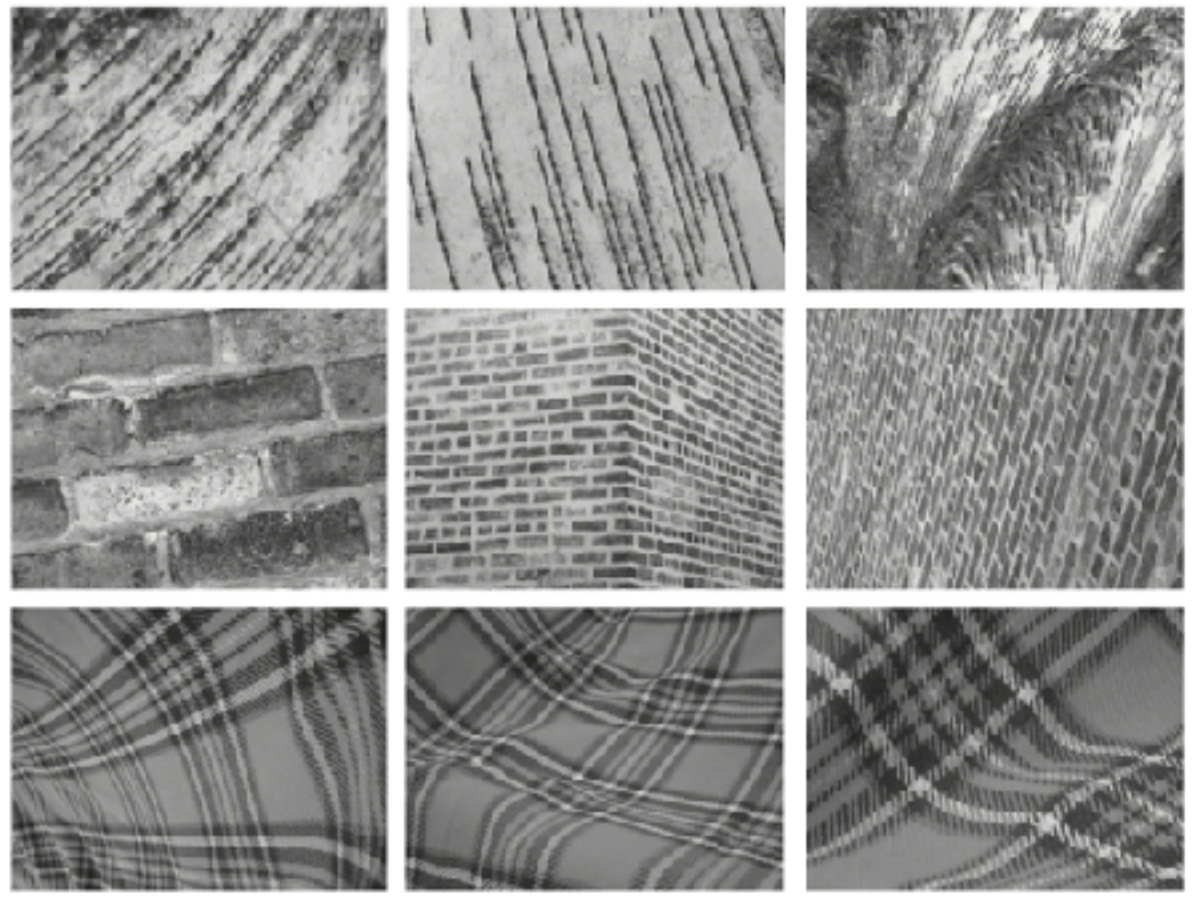
Properties:

- The filters are not learned
- Invariant to translation (+rotation)
- Stable to deformations
- Preserves energy
- Contracting

MNIST classification

Bruna & Mallat 2013

texture classification



Sifre & Mallat 2013

What can it do for scientific data analyses?

- synthesis
- parameter inference
- exploratory data analysis

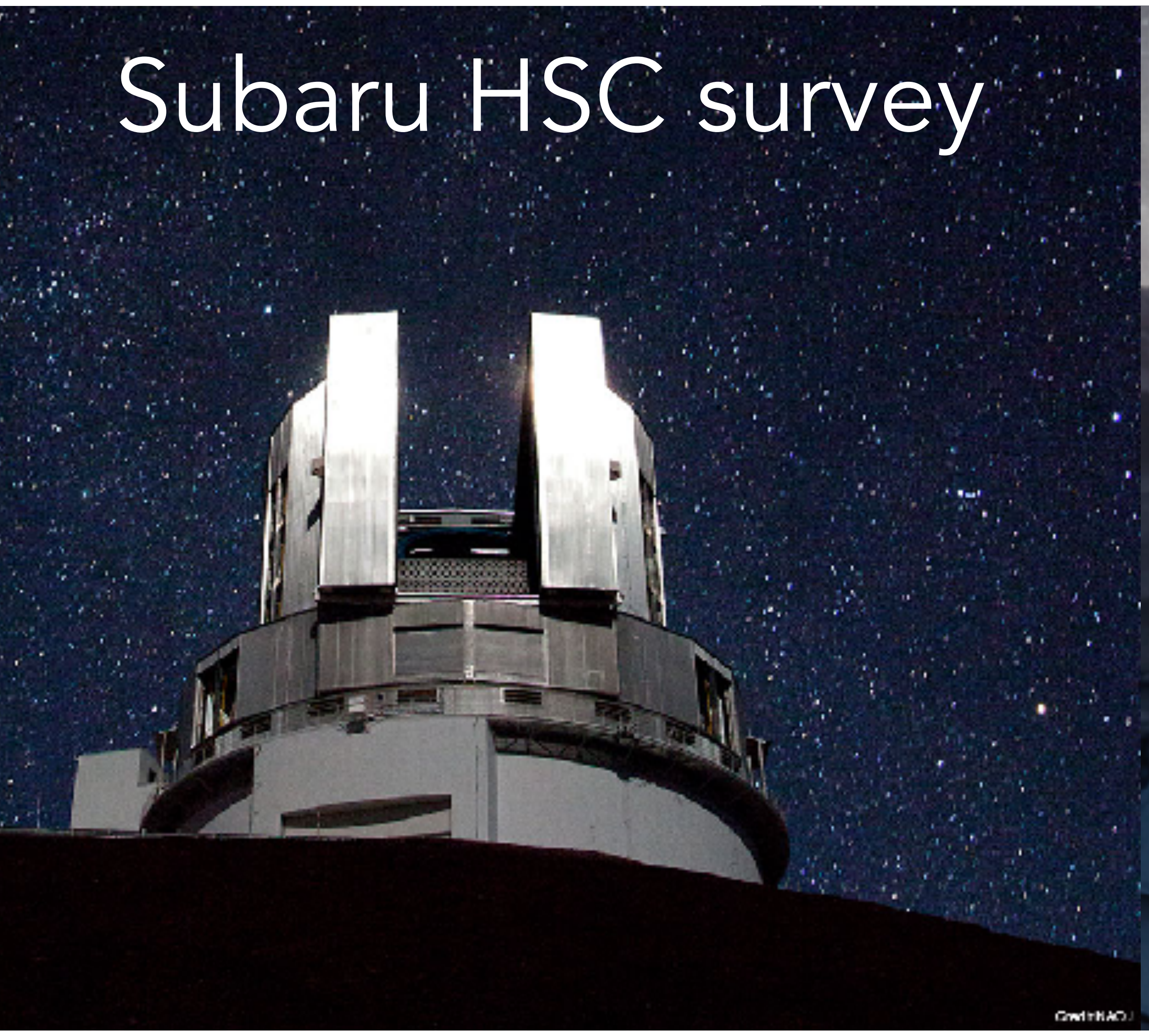
Turing pattern | Ising model | sea temperature | solar surface | cosmic matter

Syntheses with 2$^{\text{nd}}$ order scattering transform (+ min,max values)

→ for many physical fields, it captures most of the information

Cheng & Ménard, 2021

Parameter inference in cosmology — the texture of the Universe



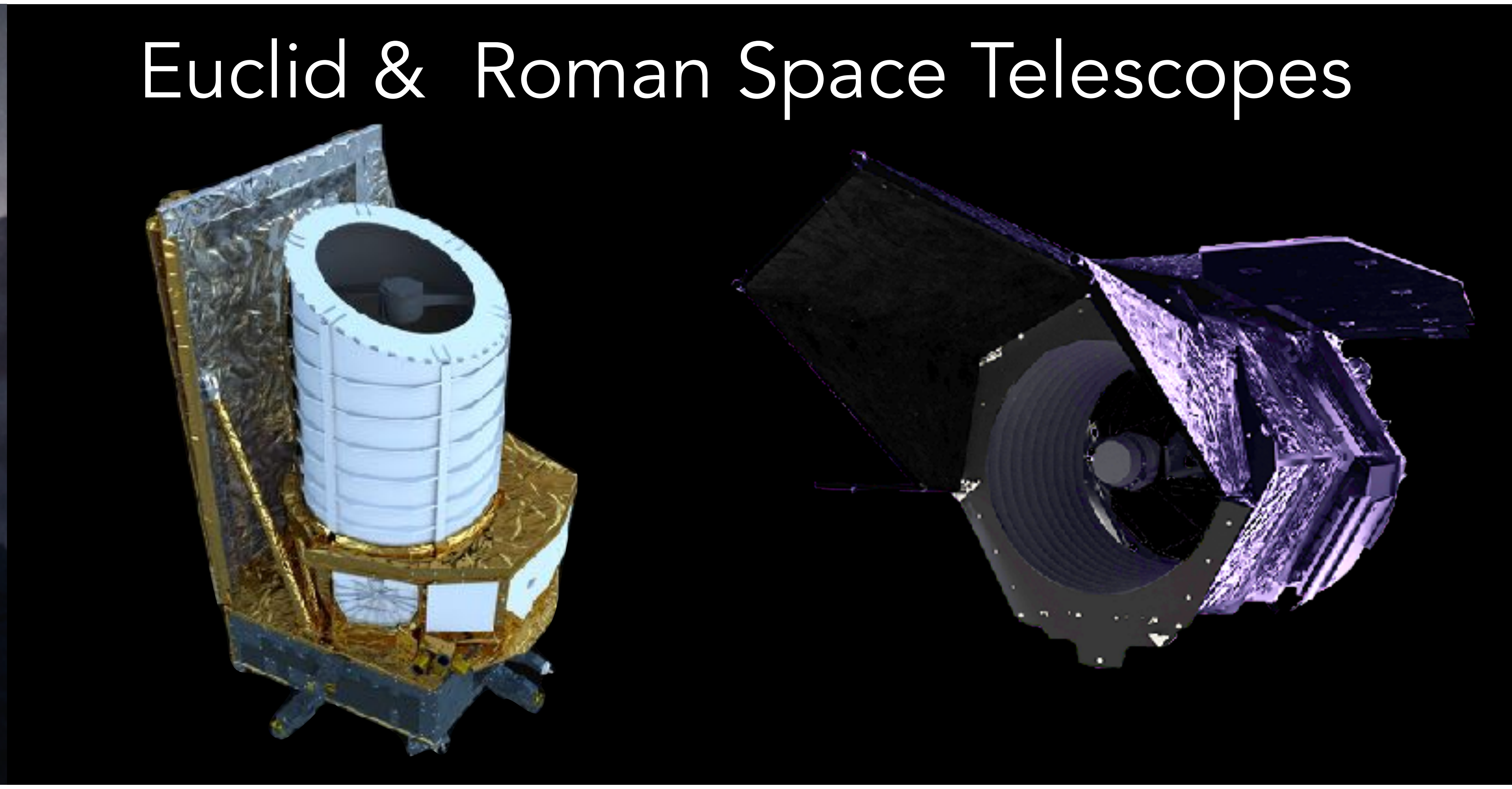
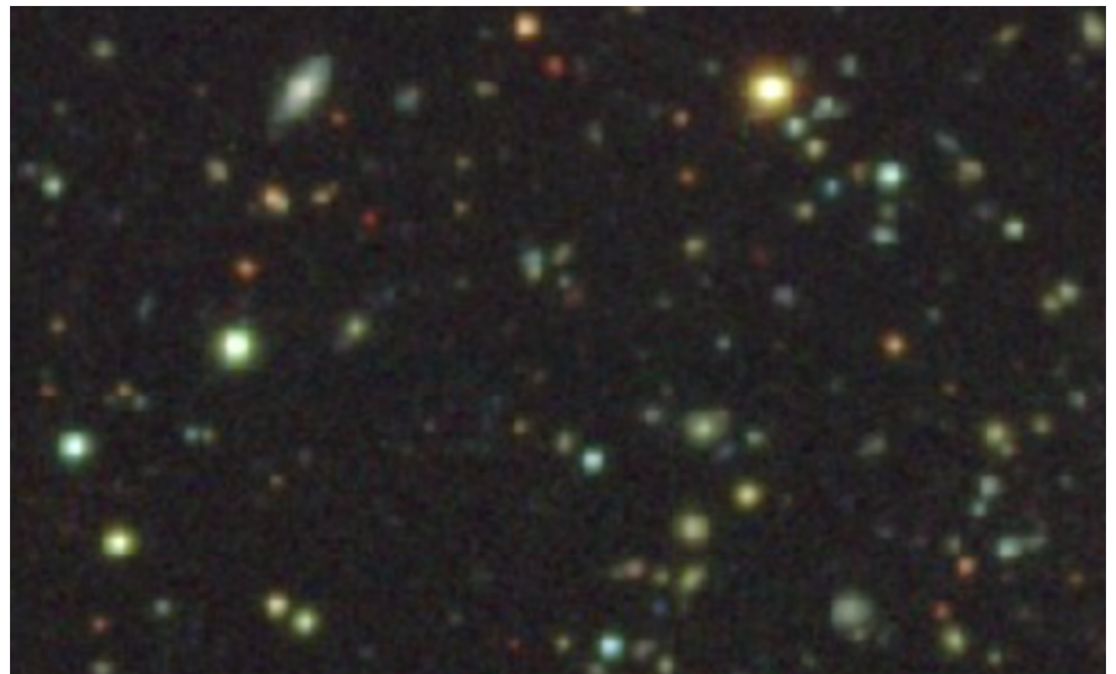
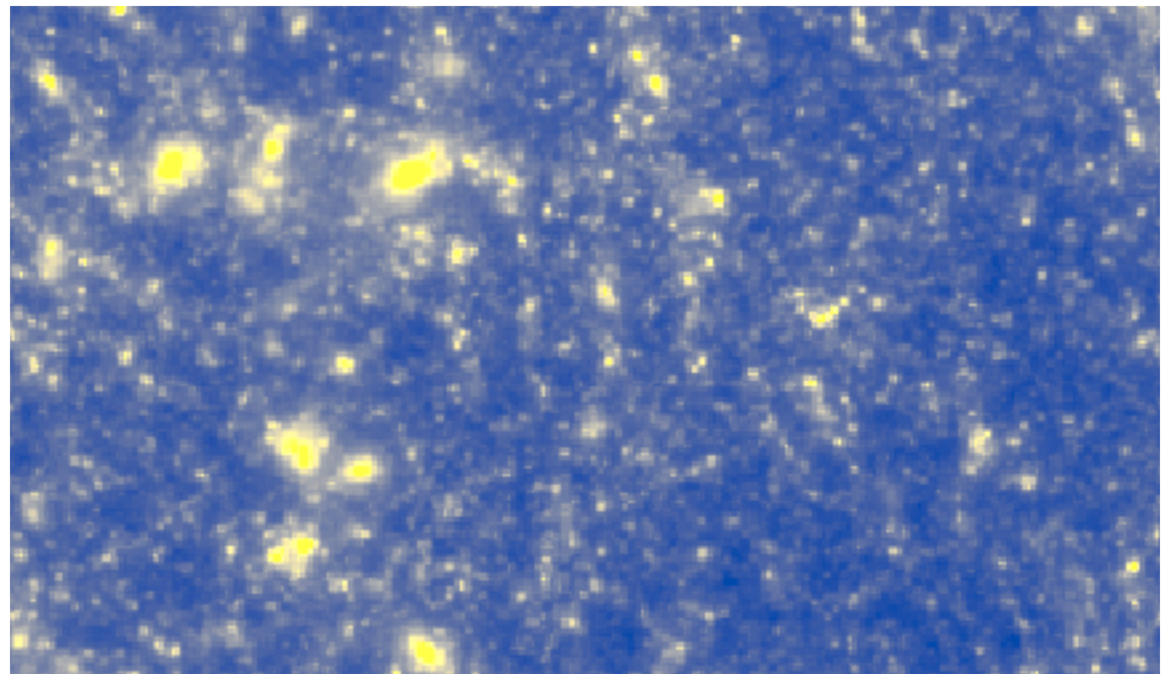


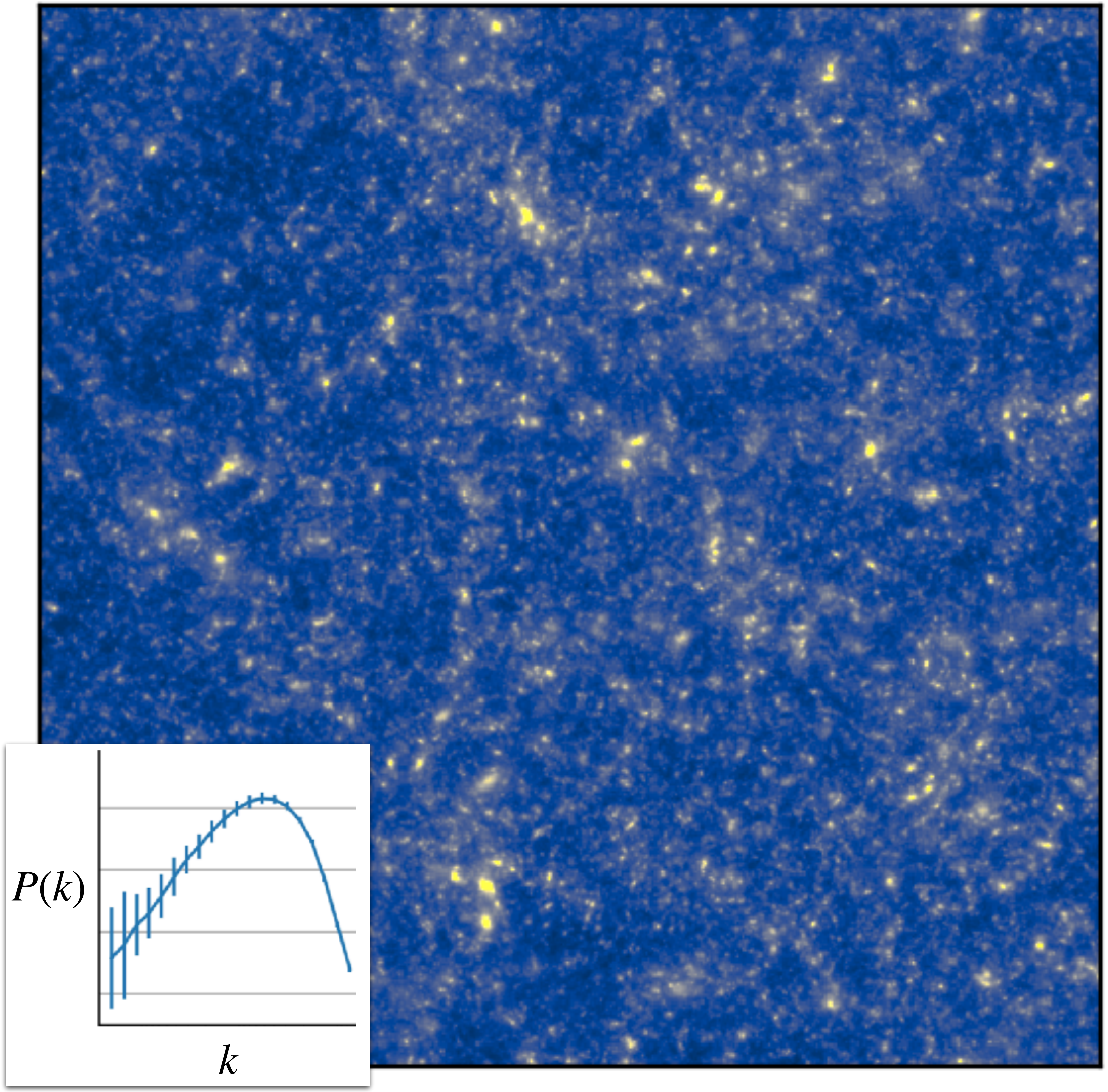
image $\rightarrow$ galaxy catalog $\rightarrow$ mass map $\rightarrow$ σ_8, Ω_m

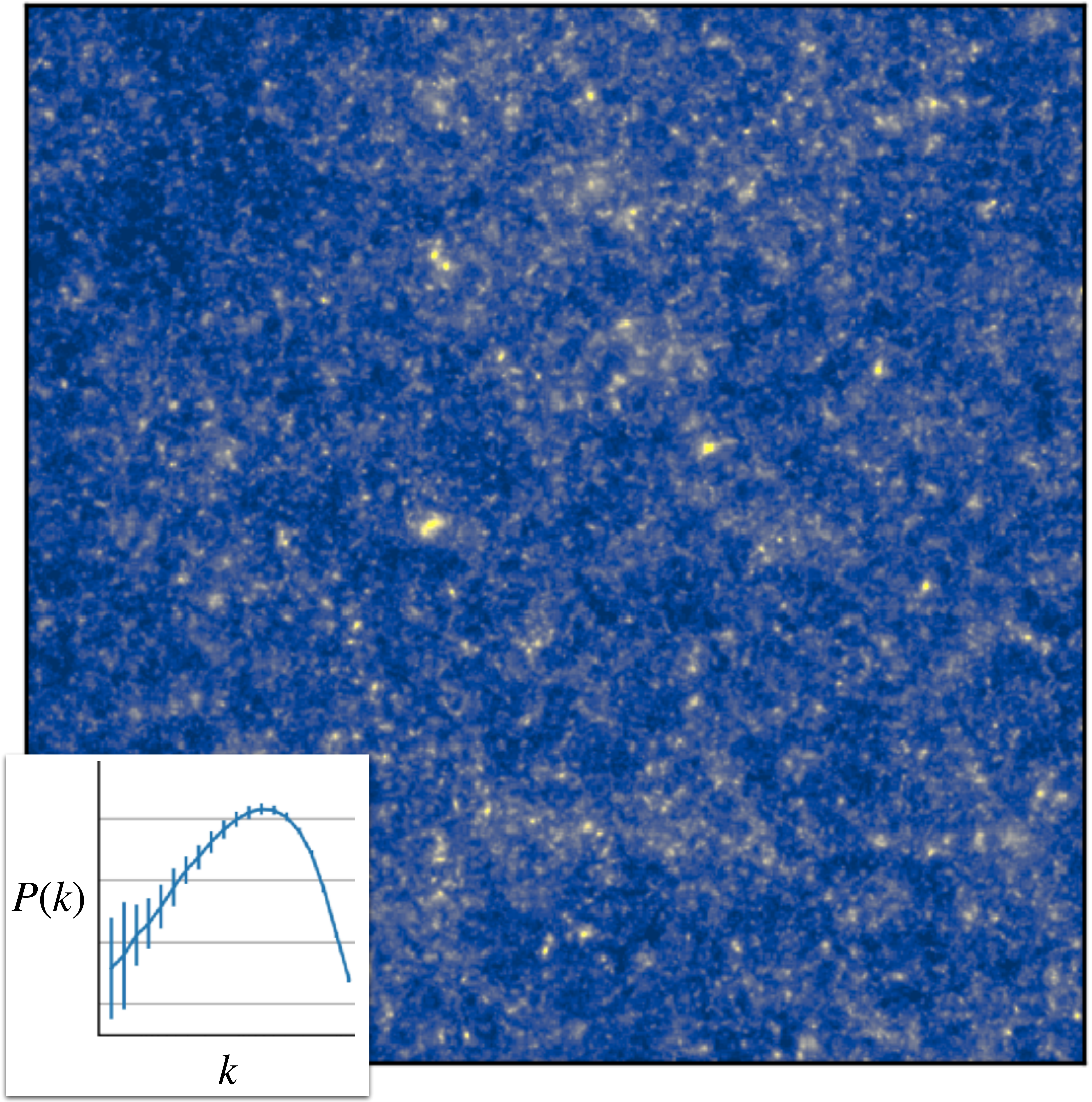


$$\begin{pmatrix} x_1, y_1, \epsilon_1 \\ x_2, y_2, \epsilon_2 \\ x_3, y_3, \epsilon_3 \\ x_4, y_4, \epsilon_4 \\ \dots \end{pmatrix}$$



Simulated weak lensing mass maps

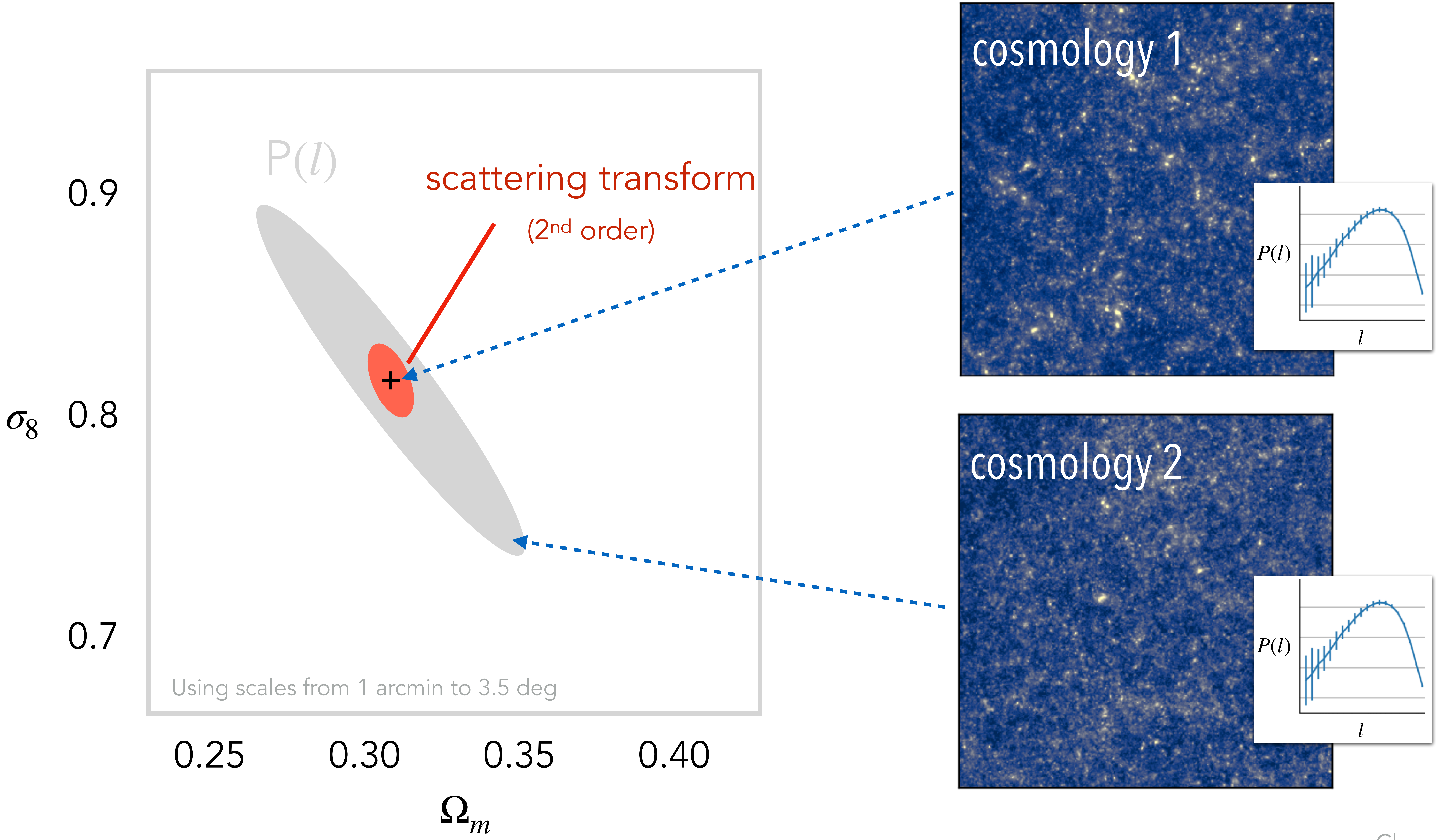




512 weak lensing maps x 100 cosmologies

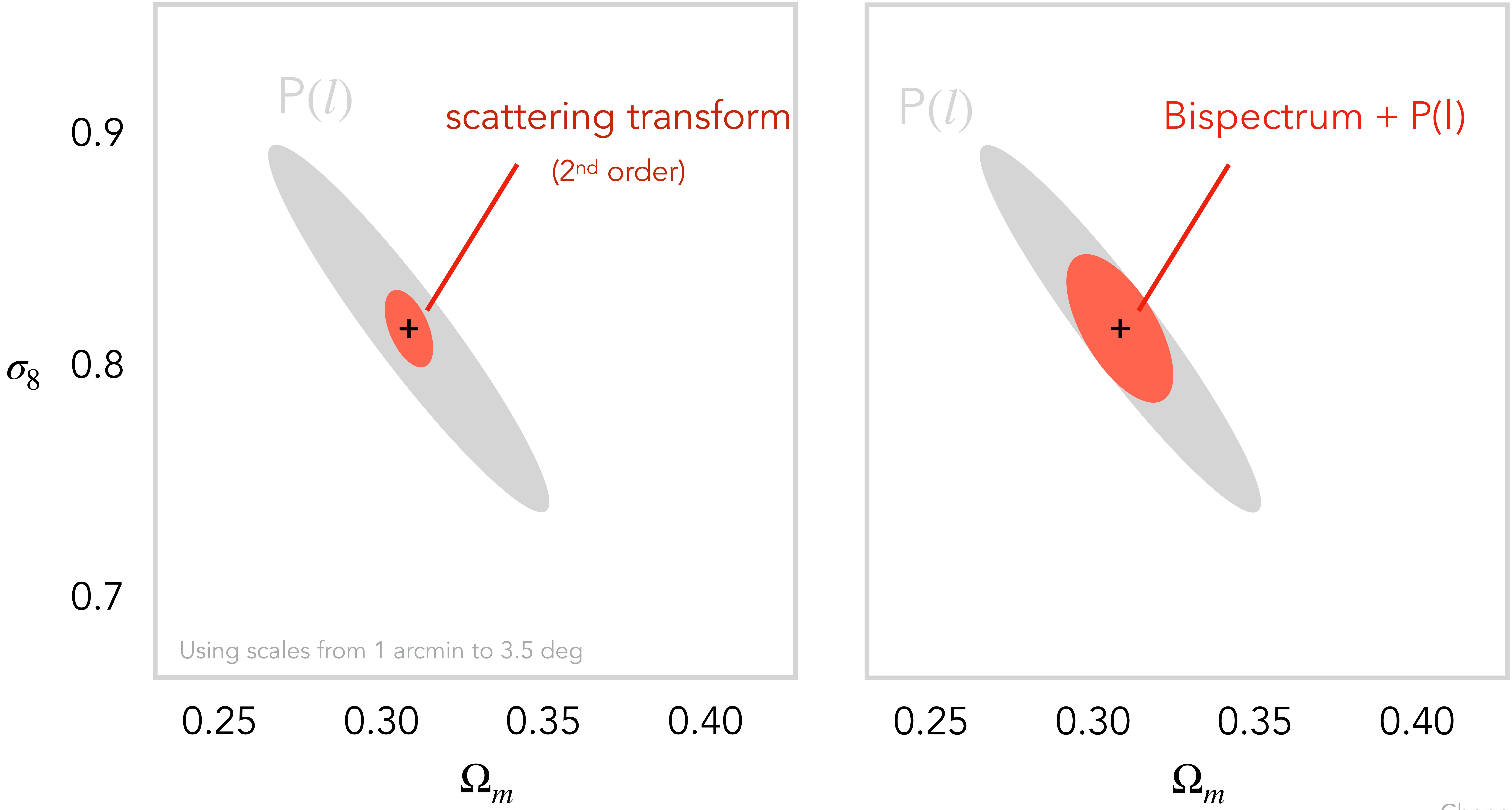
from the Columbia Univ. lensing group. See Matilla Zorrilla et al. 2016, Gupta et al. 2018

Scattering coefficients vs. power spectrum



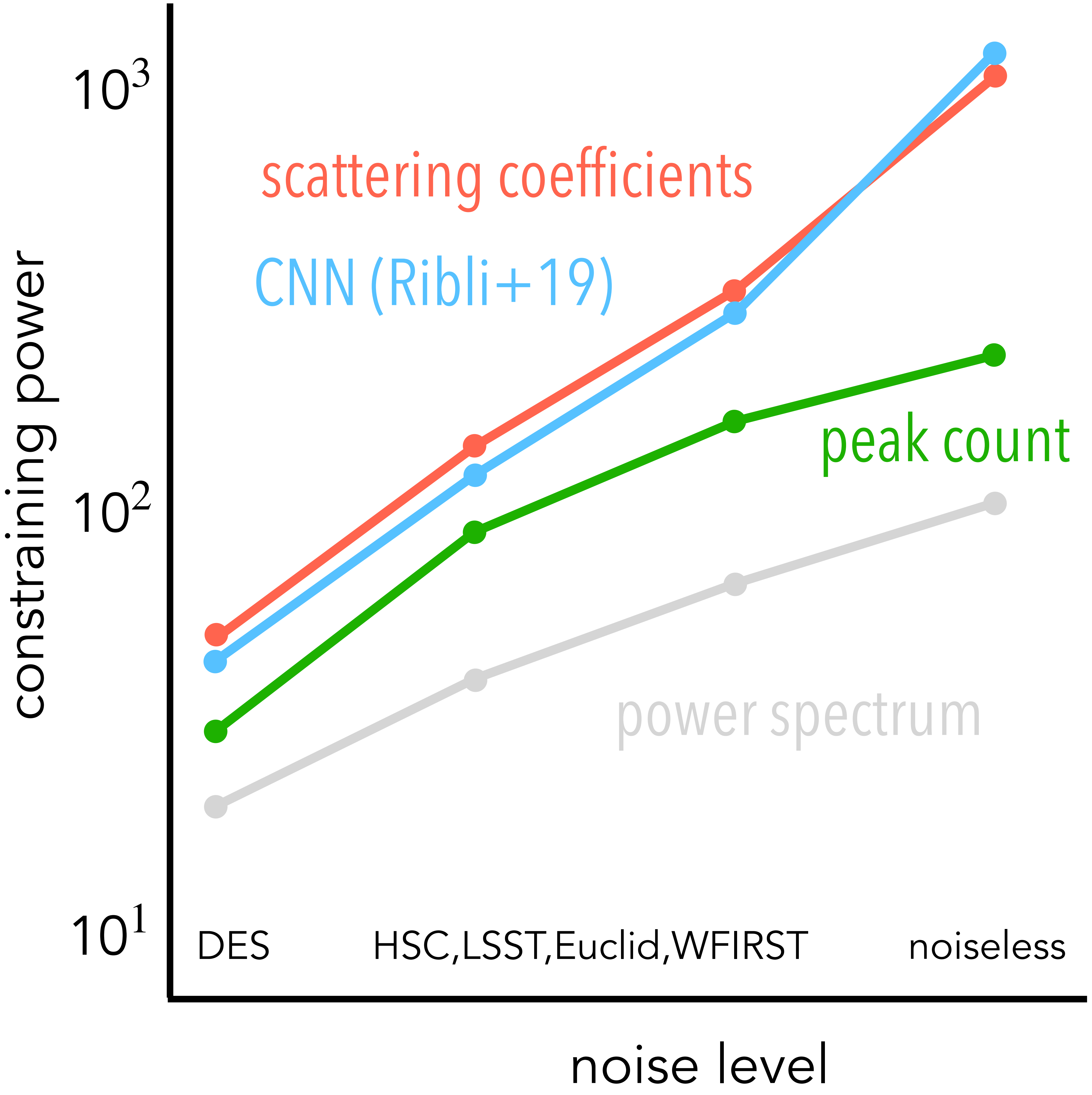
Cheng & BM, 2021

Scattering coefficients vs. bispectrum



Cheng & BM, 2021

Scattering transform performance with noise



How many coefficients were used?

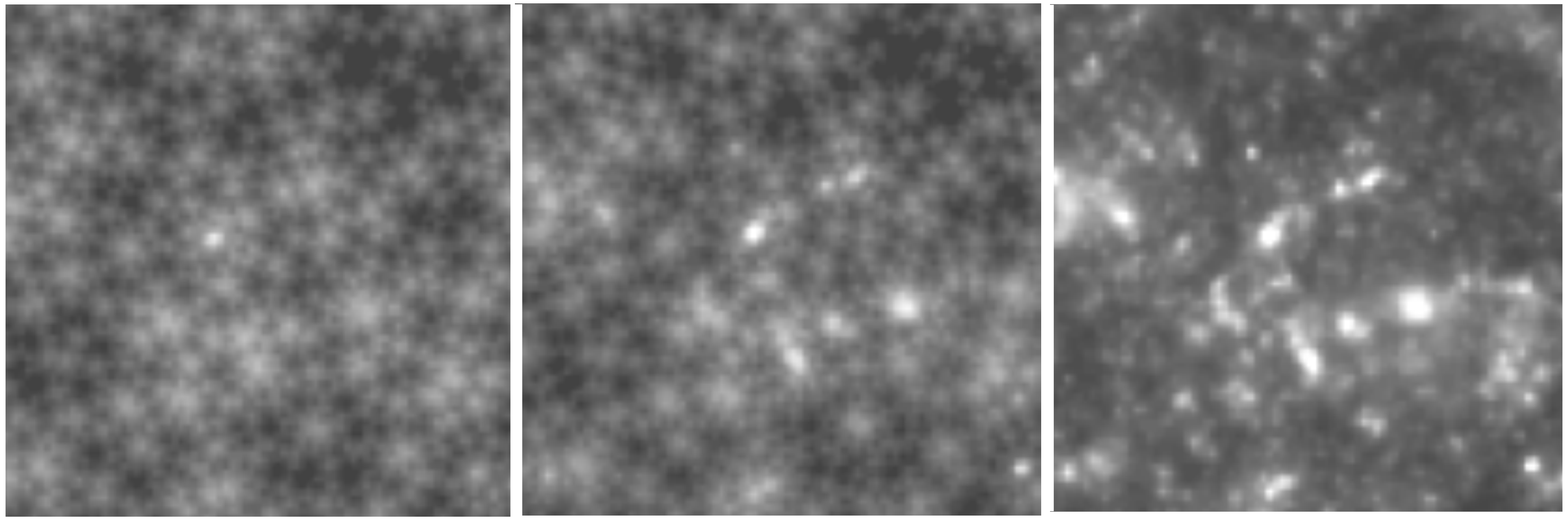
$P(k)$: 20

scattering transform: 37

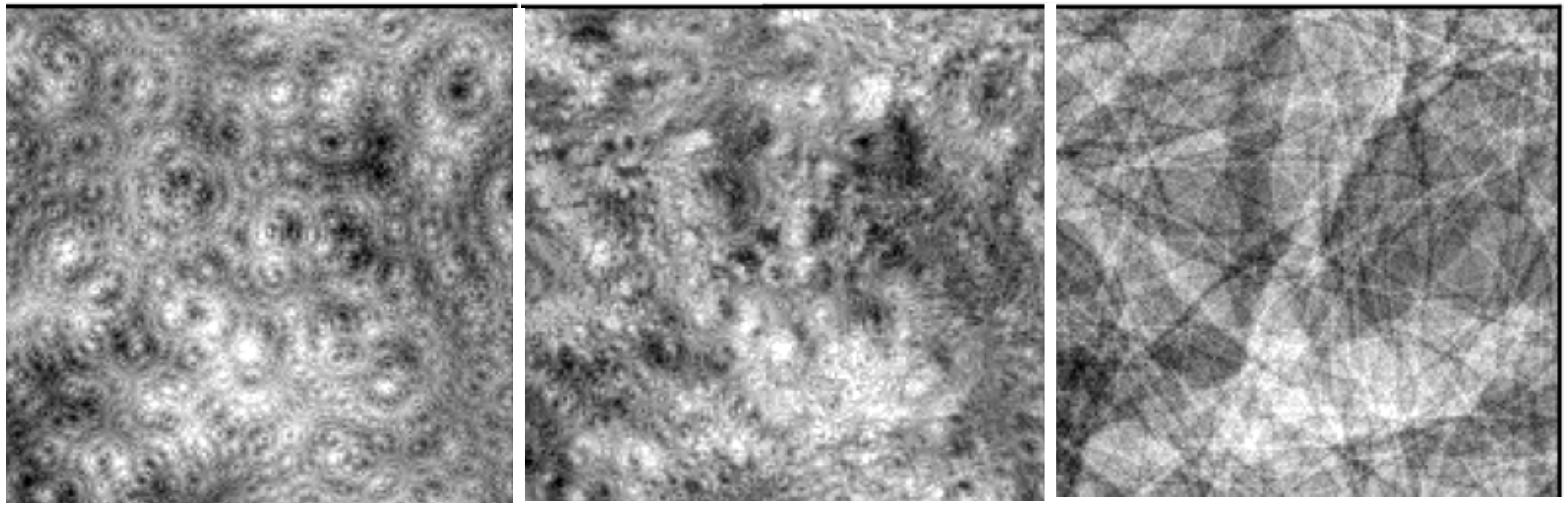
CNN: millions

Cheng & BM, 2021

Scattering transform: interpretability



structure sparsity $s_{21} \equiv S_2 / S_1$



structure shape $s_{22} \equiv S_2^{\parallel} / S_2^{\perp}$

How many coefficients were used?

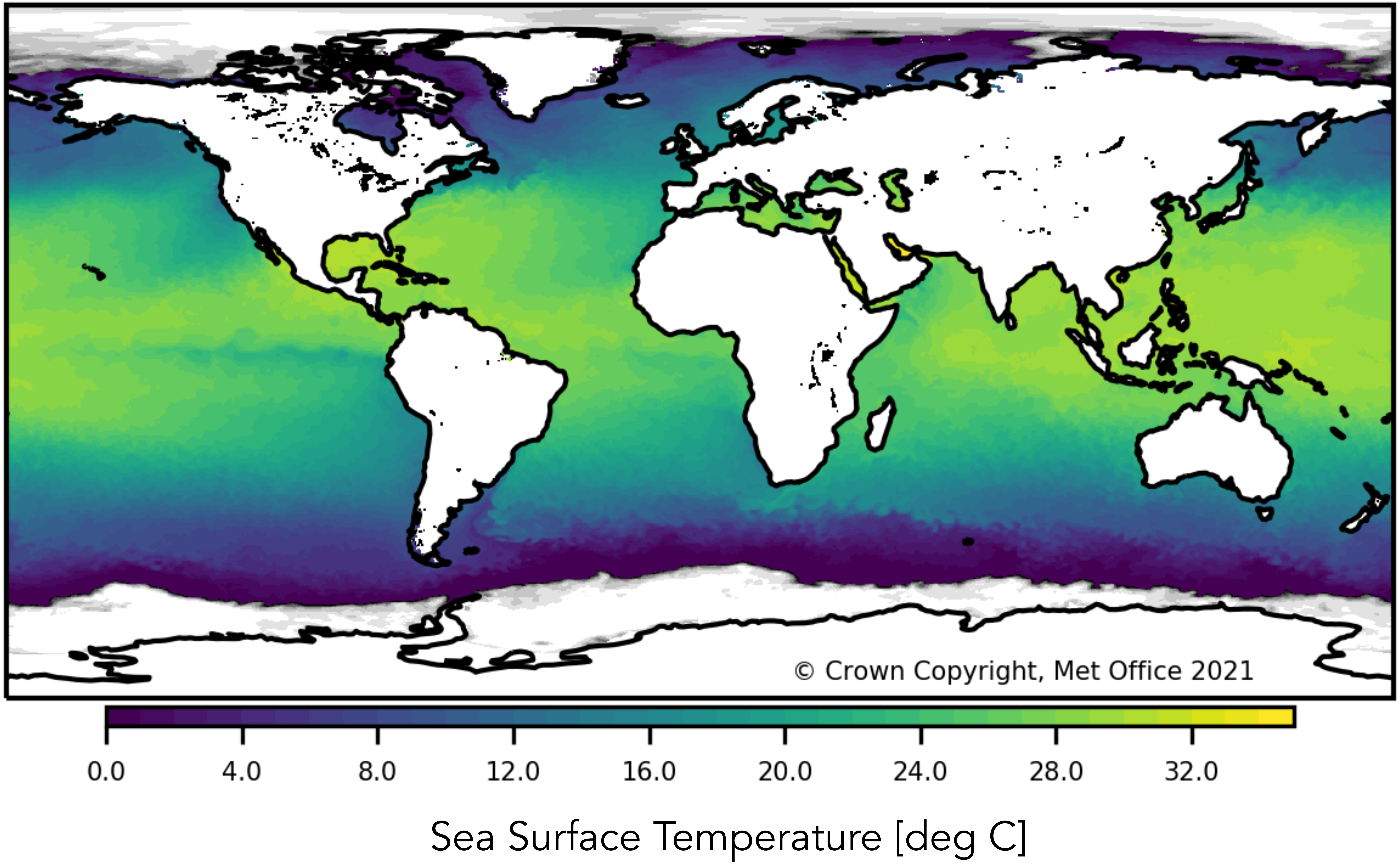
$P(k)$: 20

scattering transform: 37

CNN: millions

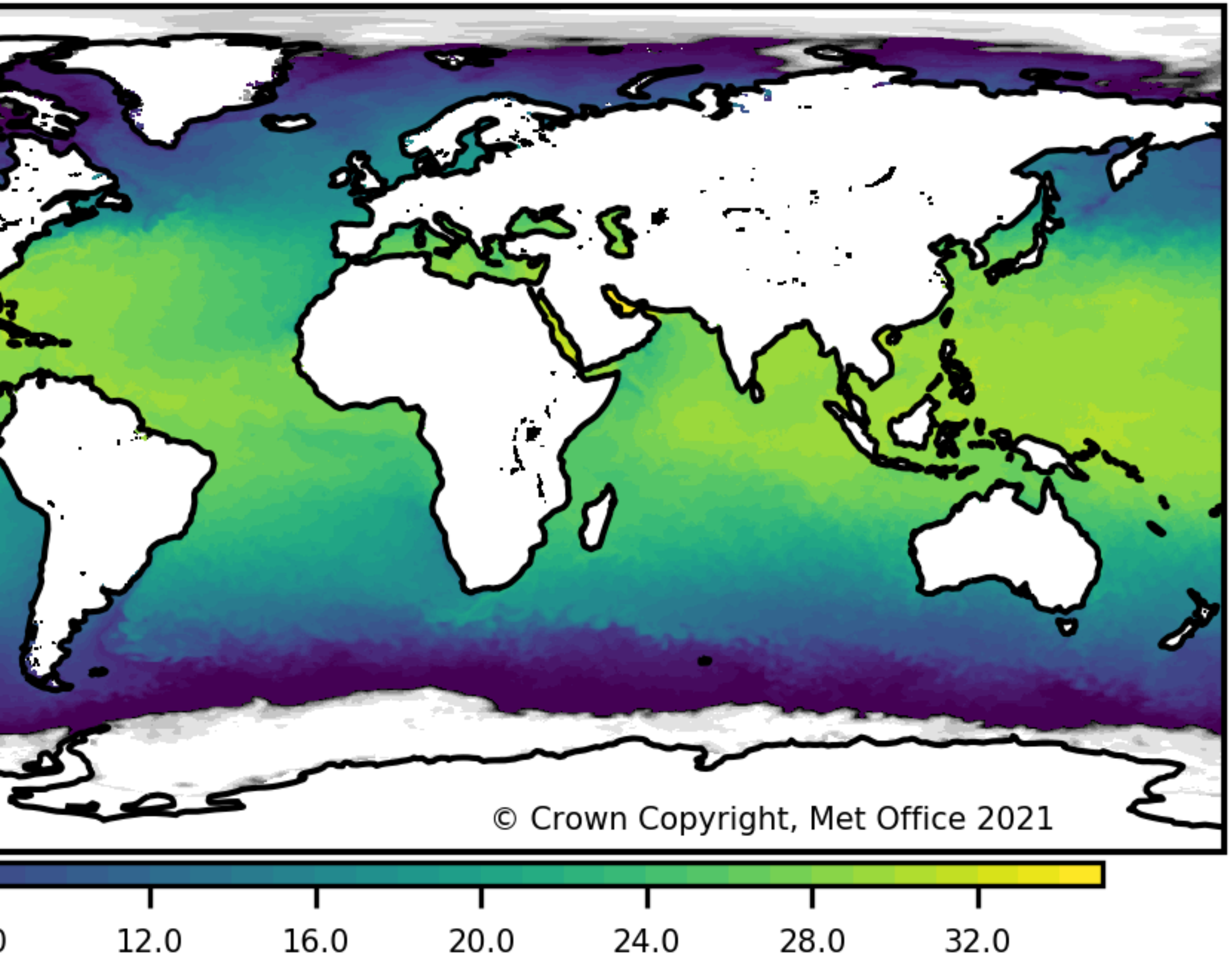
Cheng & BM, 2021

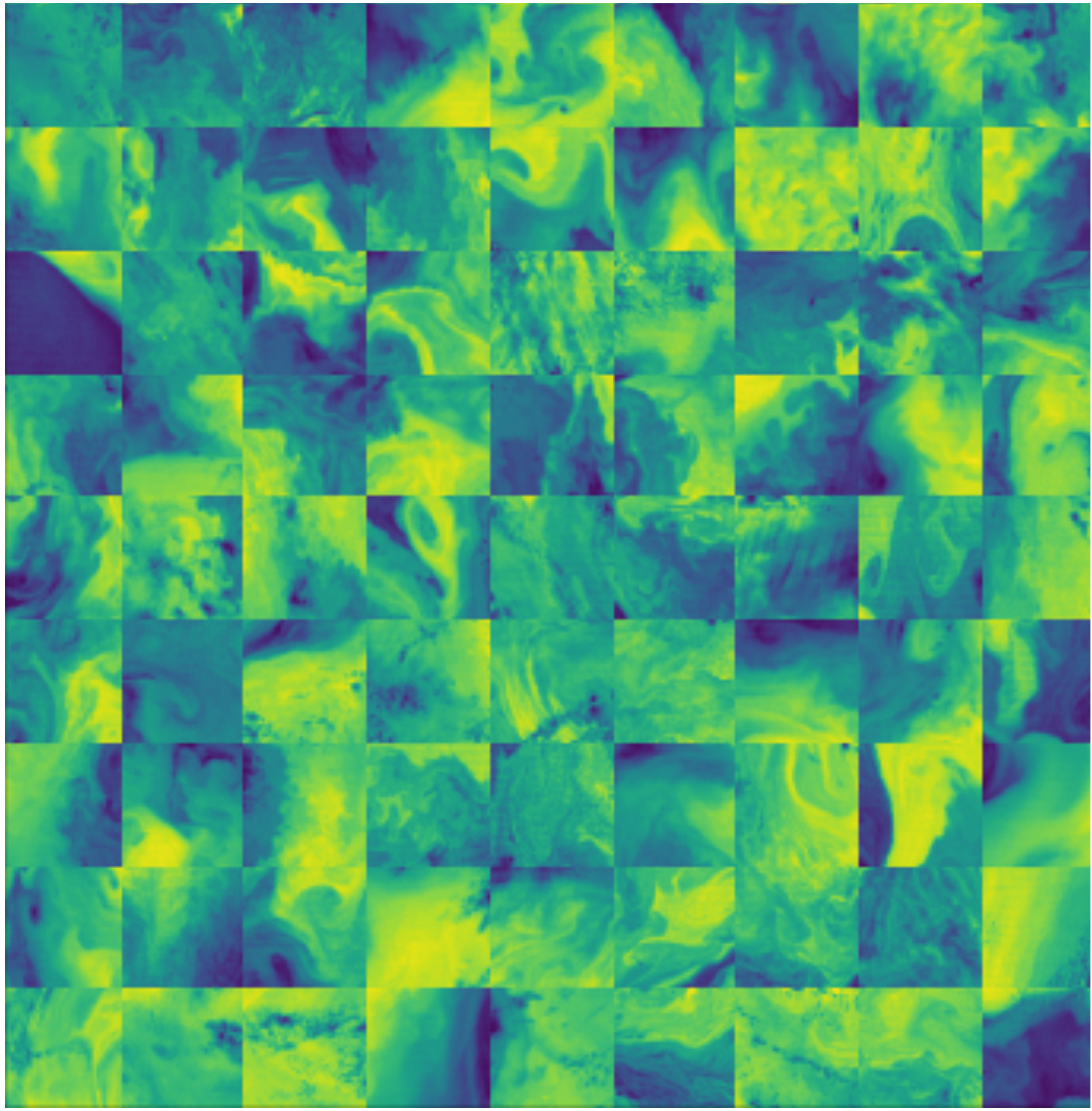
Exploratory data analysis



CNN analysis by Prochaska, Cornillon, Reiman (2021)

Exploratory data analysis

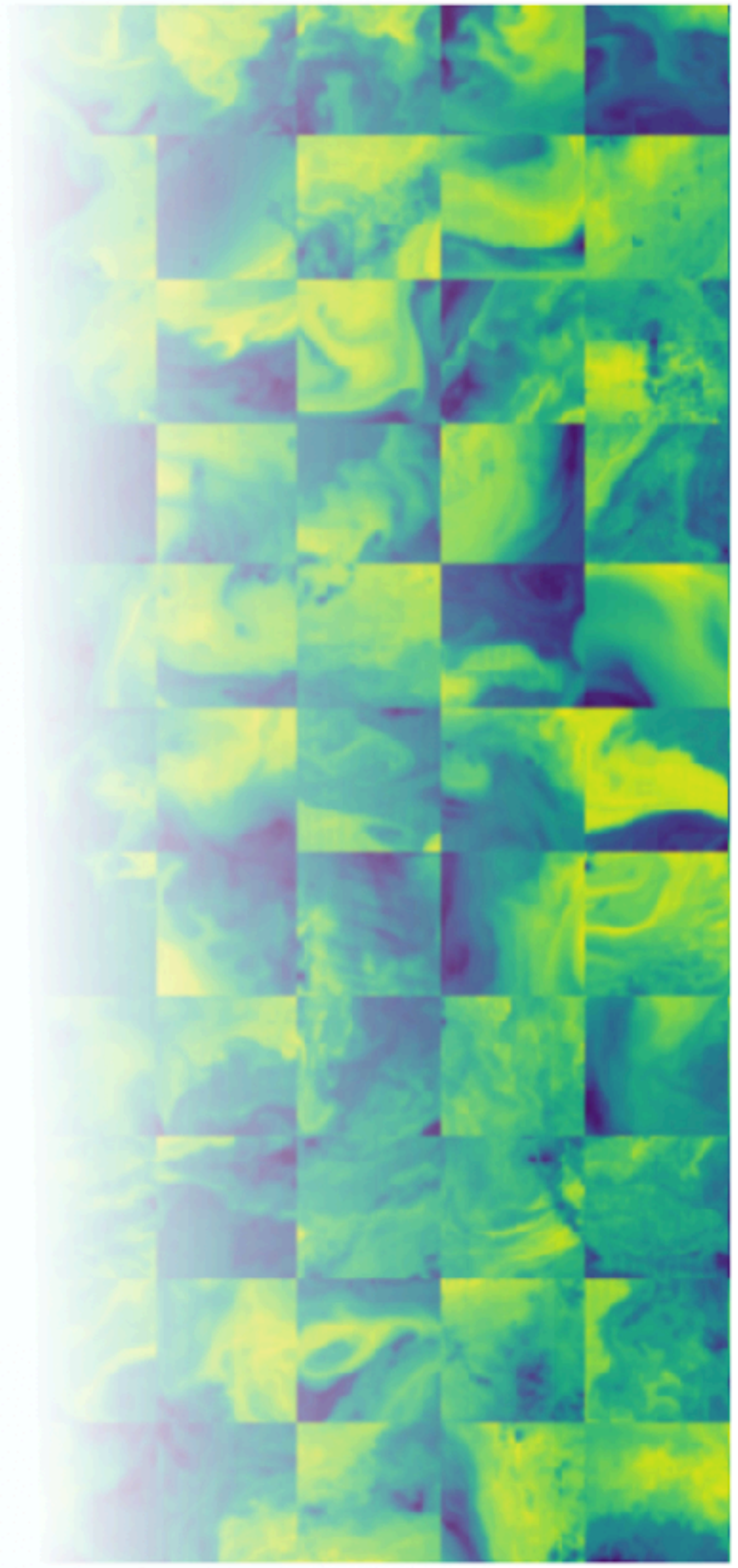


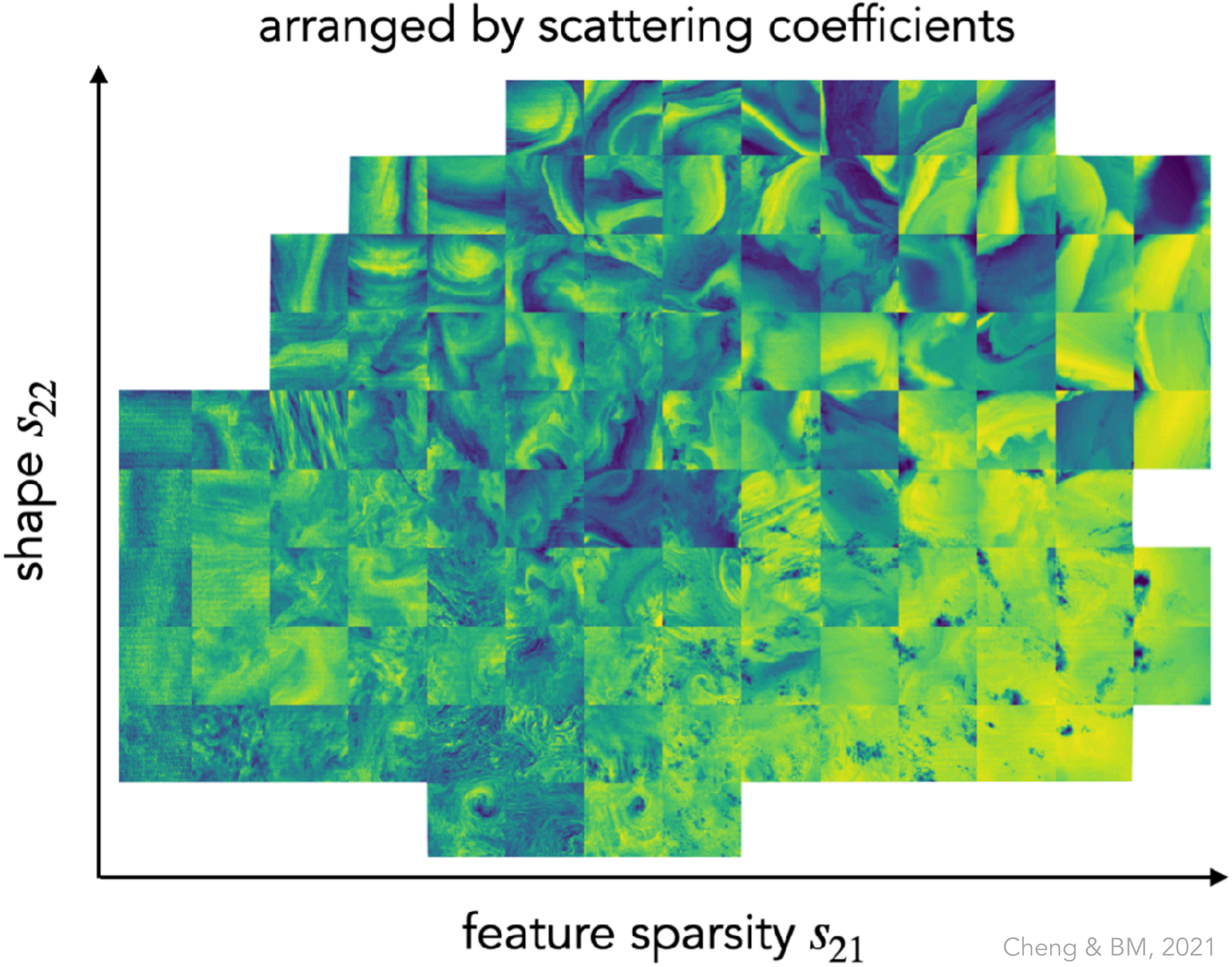


Sea Surface Temperature [deg C]

CNN analysis by Prochaska, Cornillon, Reiman (2021)

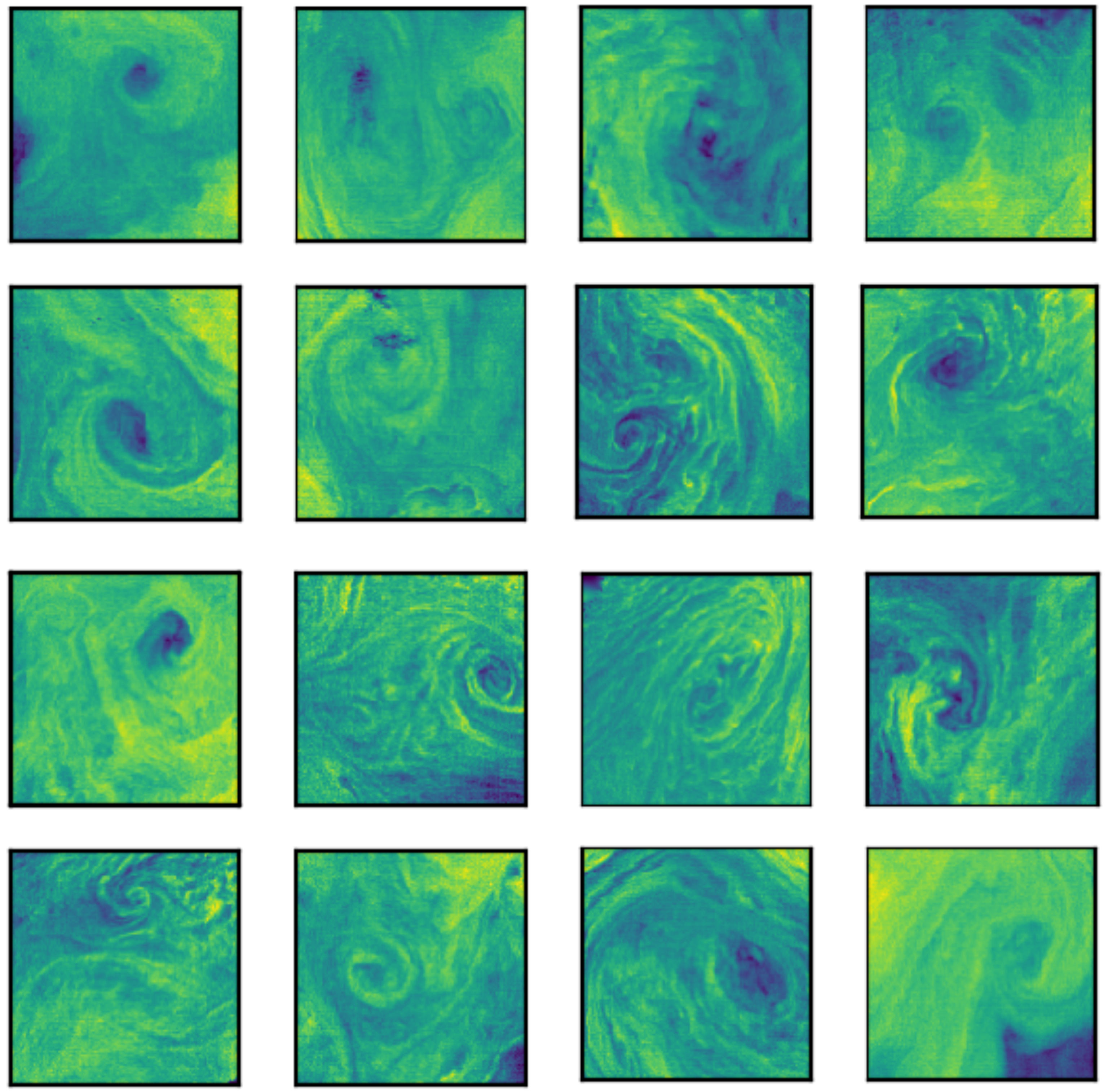
Exploratory data analysis

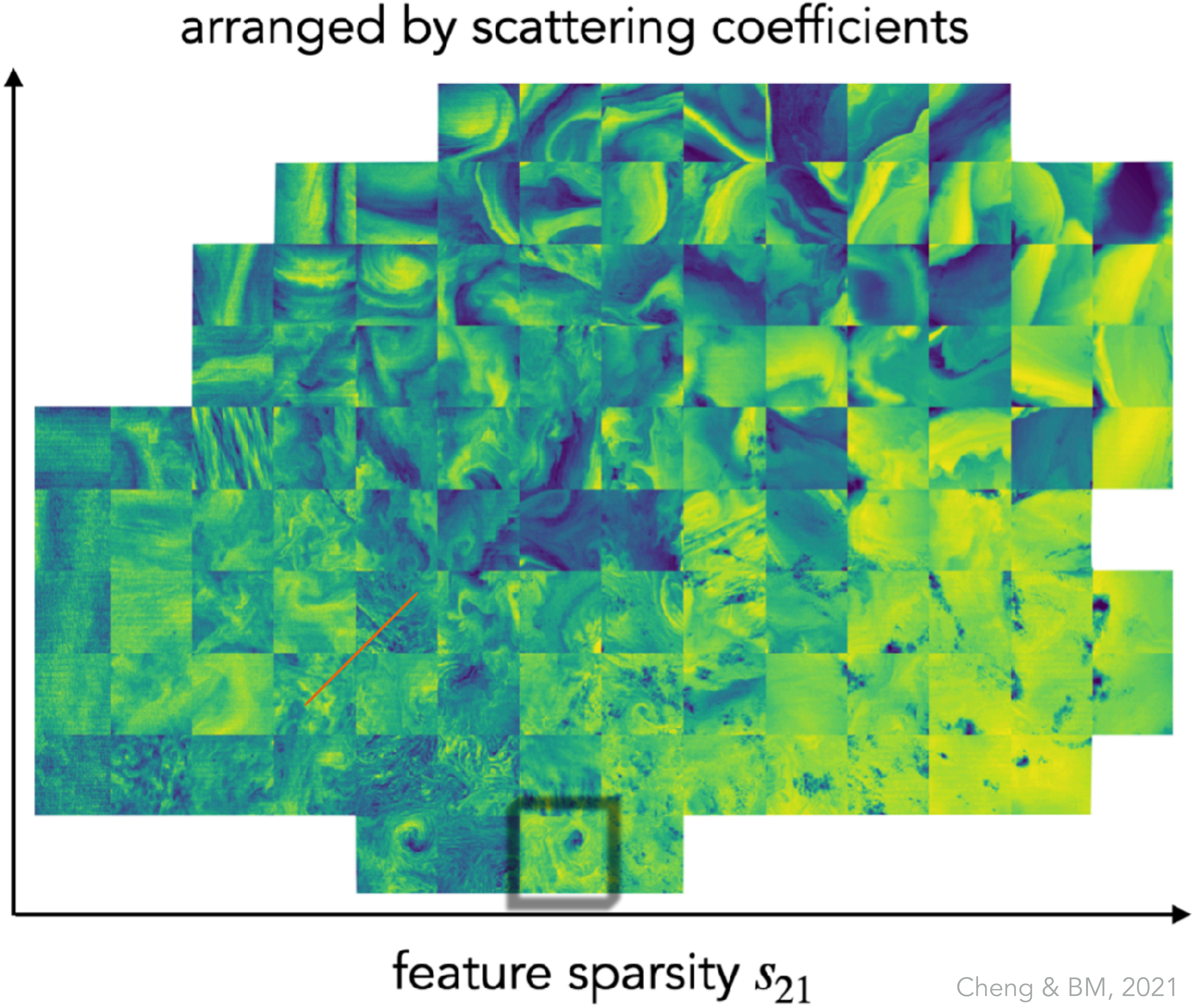




Cheng & BM, 2021

Exploratory data analysis





Cheng & BM, 2021

Exploratory data analysis: objects

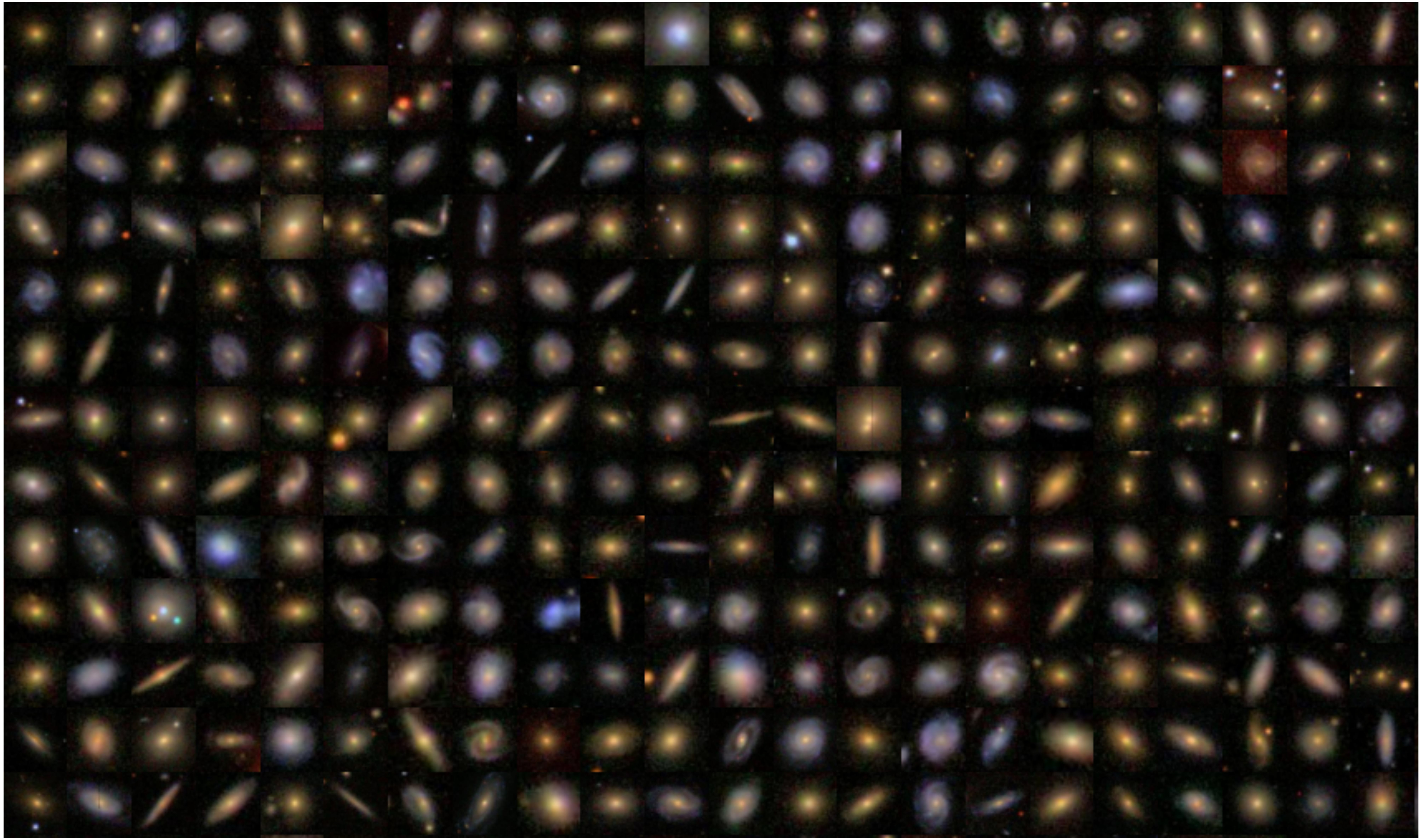


image credit: Sloan Digital Sky Survey

Exploratory data analysis: objects

Cheng & BM, 2021

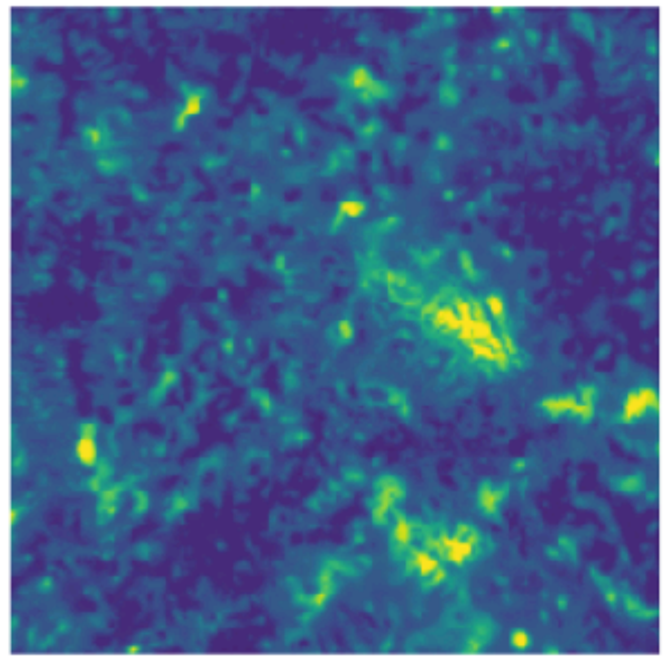
“mathematical” neural networks

- ✓ texture classification
- ✓ synthesis of physical fields
- ✓ parameter inference
- ✓ exploratory data analysis

A guide to the Scattering Transform

Cheng & Ménard (2021) arXiv:2112.01288

10 — 100 — 1,000 — 10,000 — 100,000 — 1 million



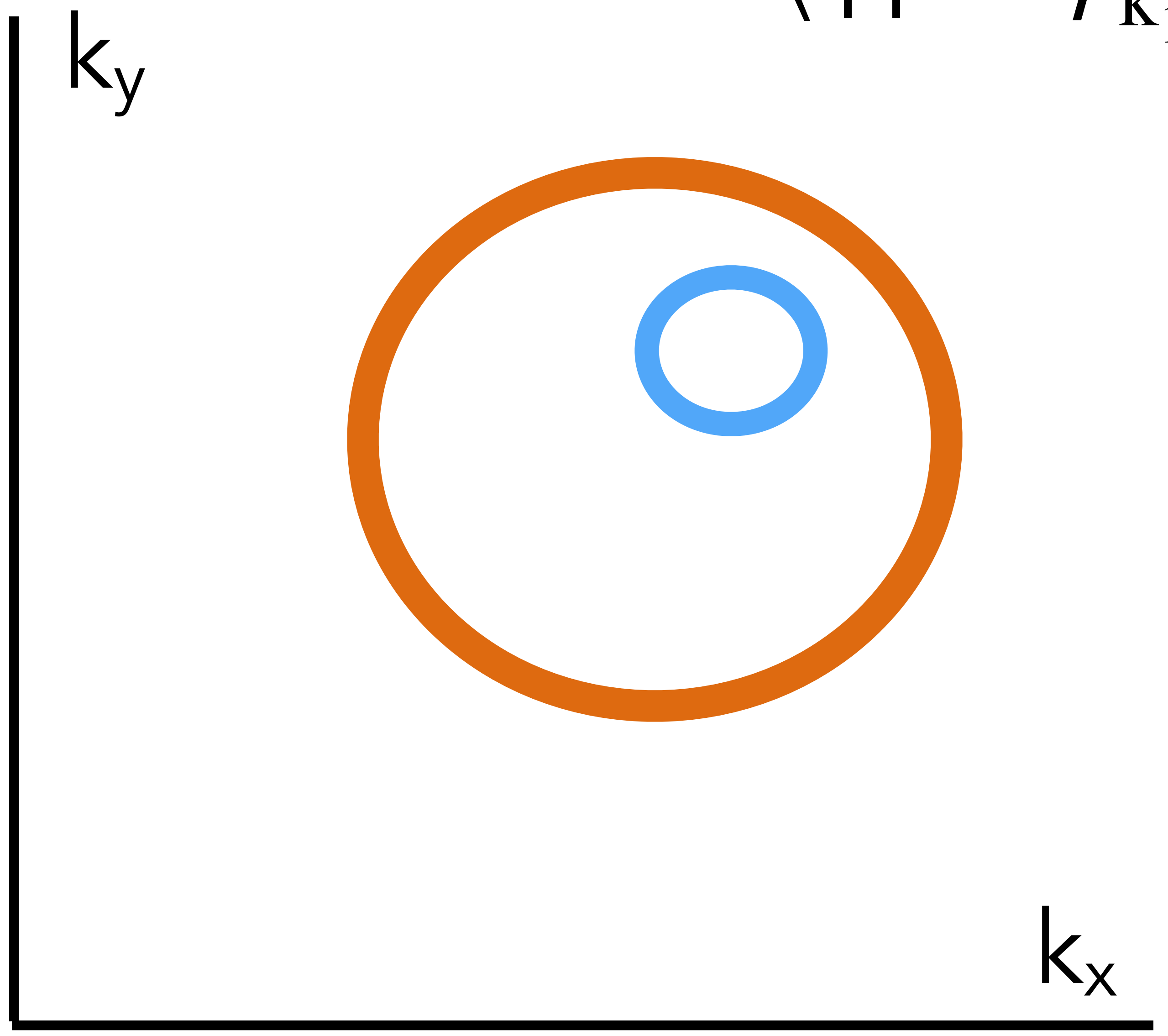
scattering transform

scattering covariance

phase harmonic transform

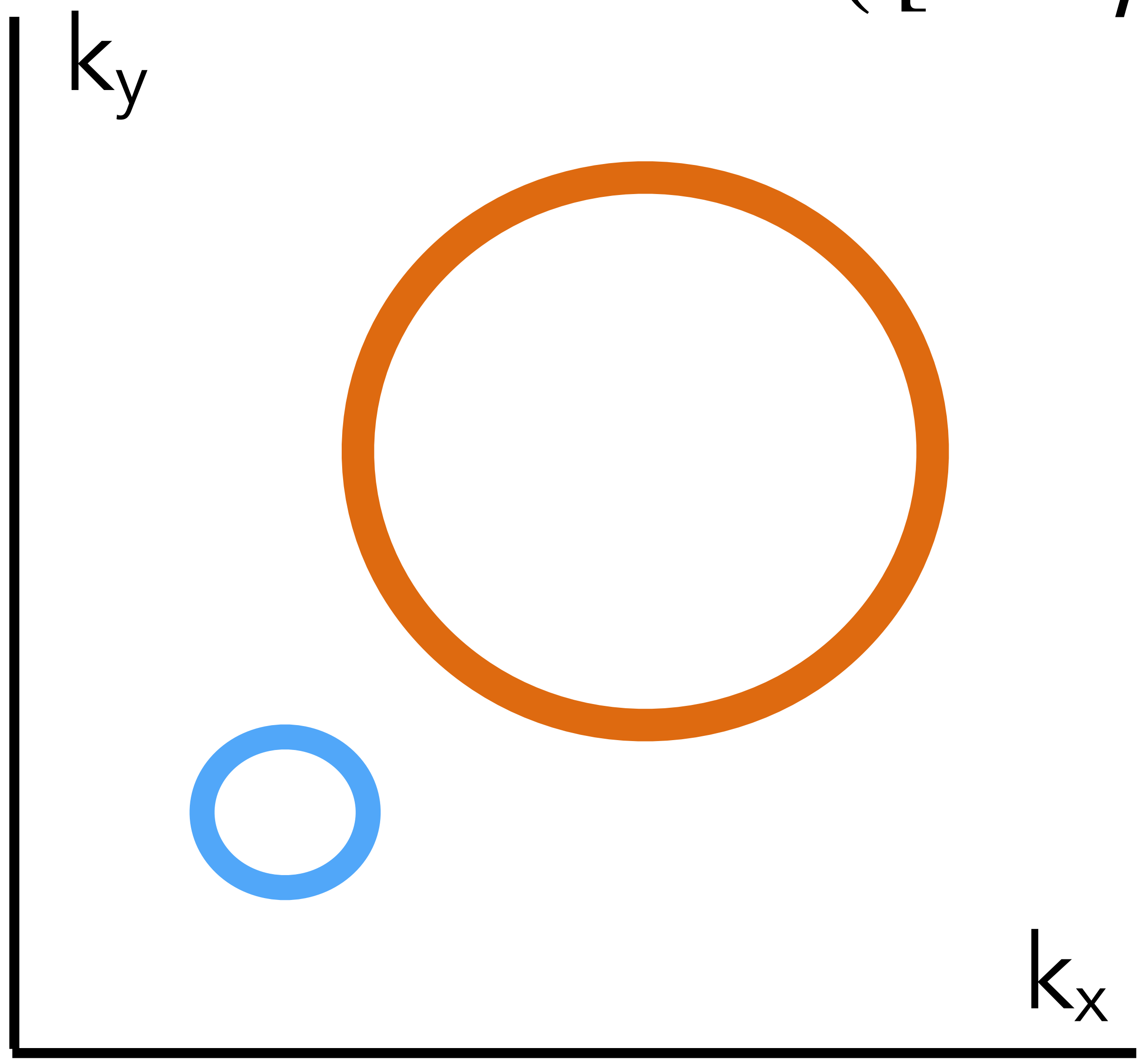
What limits the scattering transform?

$$\langle \, | \, | I * \psi_{k_1} | * \psi_{k_2} | \, \rangle$$

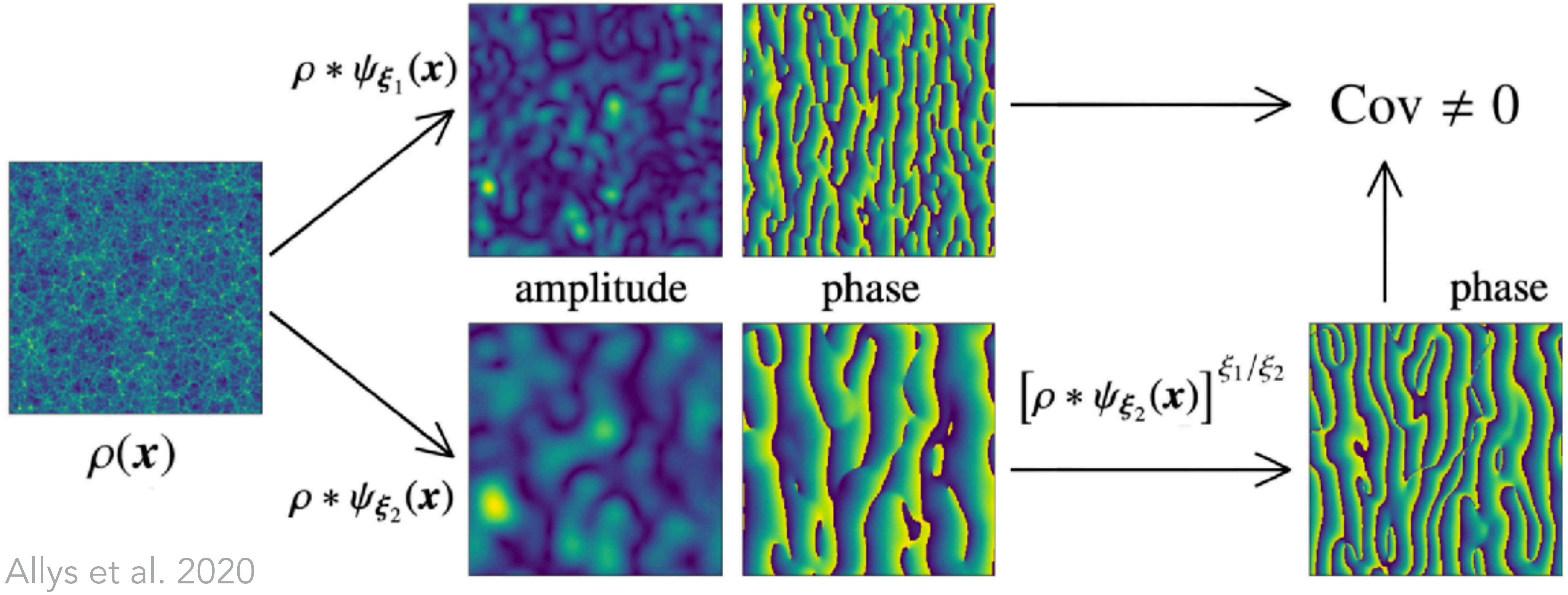


phase harmonic transform

$$\mathrm{covar}(\,[\mathrm{I} \star \psi_1]^{\mathrm{p}} \, , [\mathrm{I} \star \psi_2]^{\mathrm{q}}\,)$$



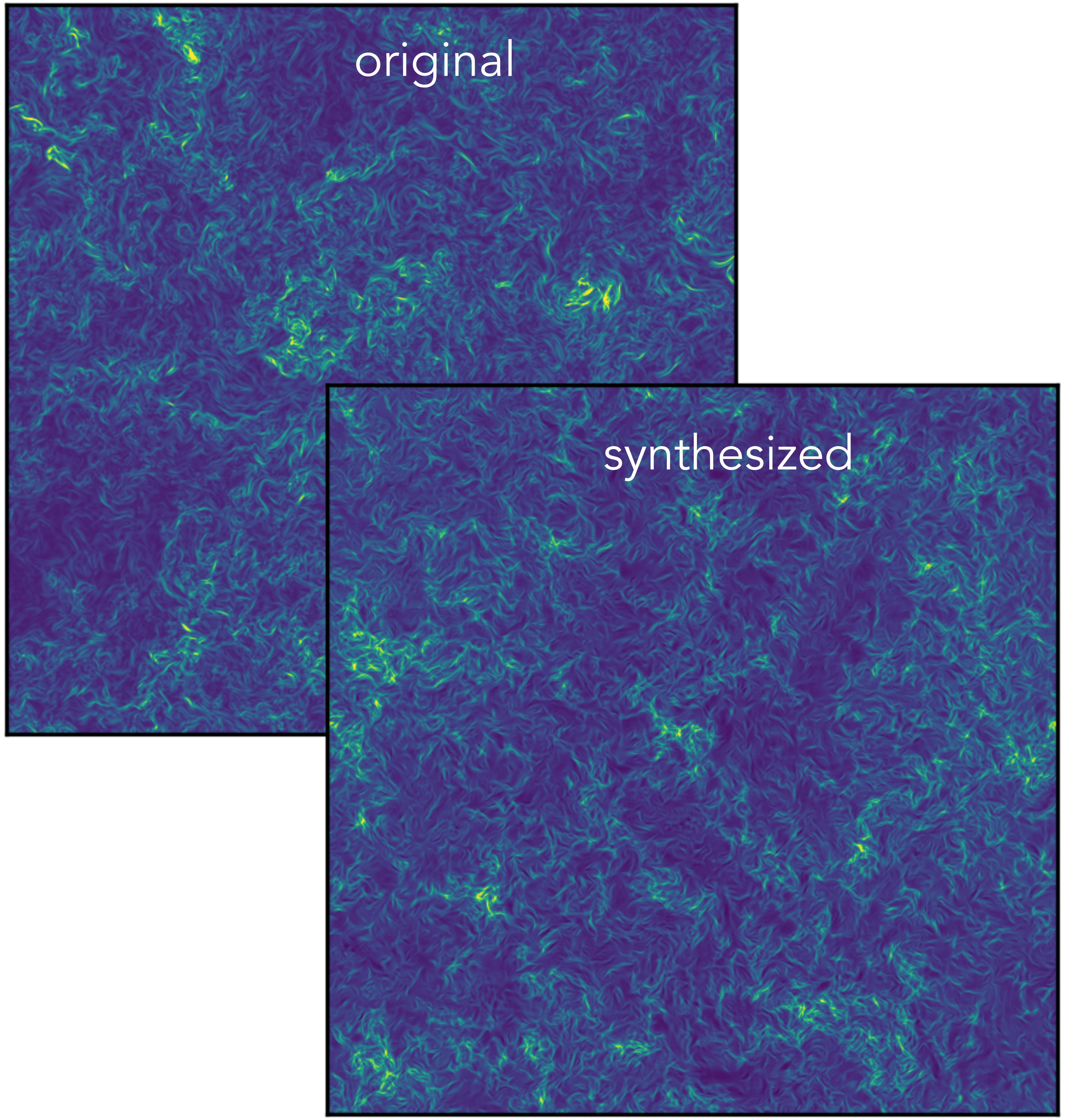
Mallat, Zhang & Rochette 2018
Zhang & Mallat 2019

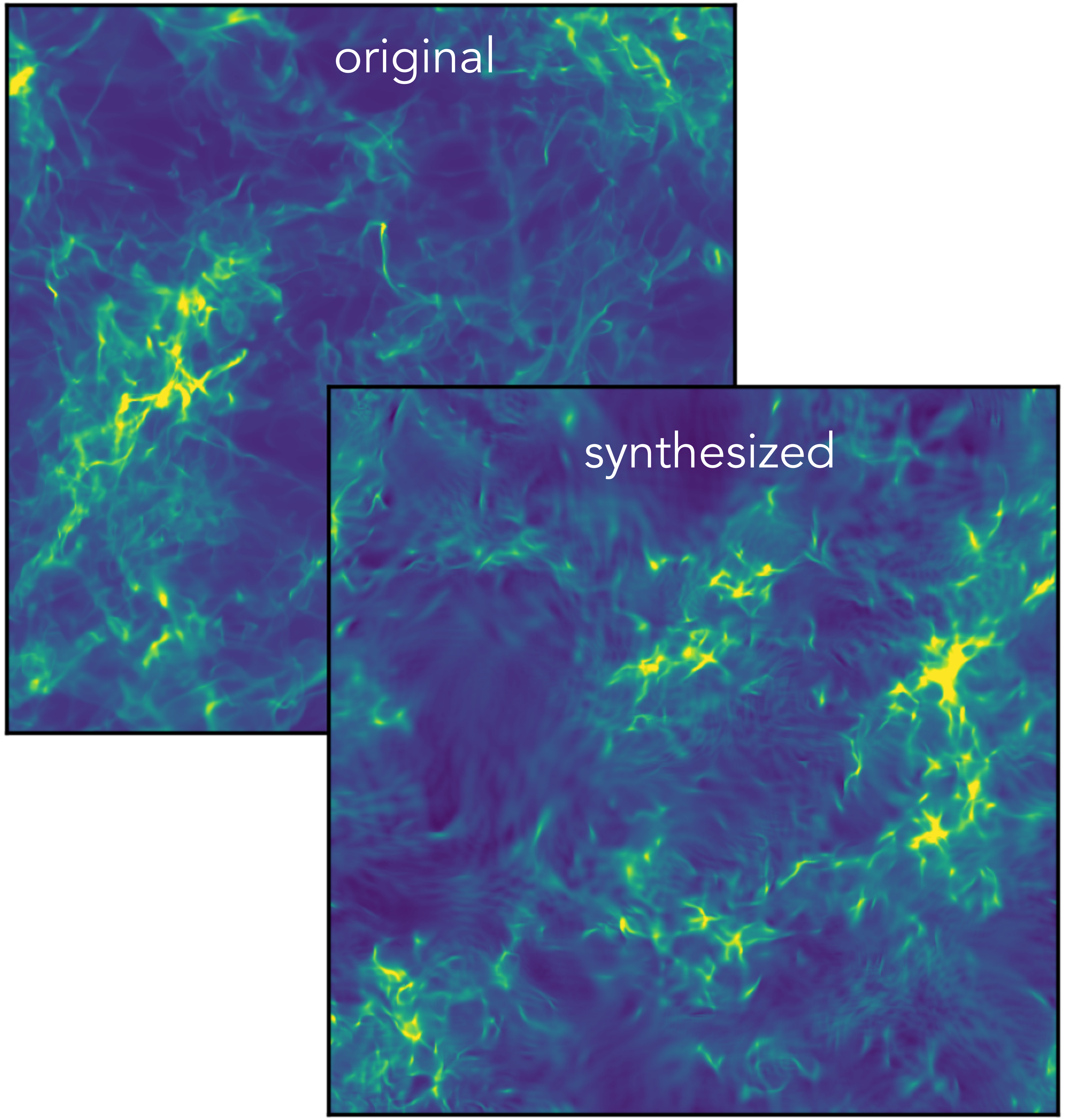


Allys et al. 2020

scattering covariance estimates

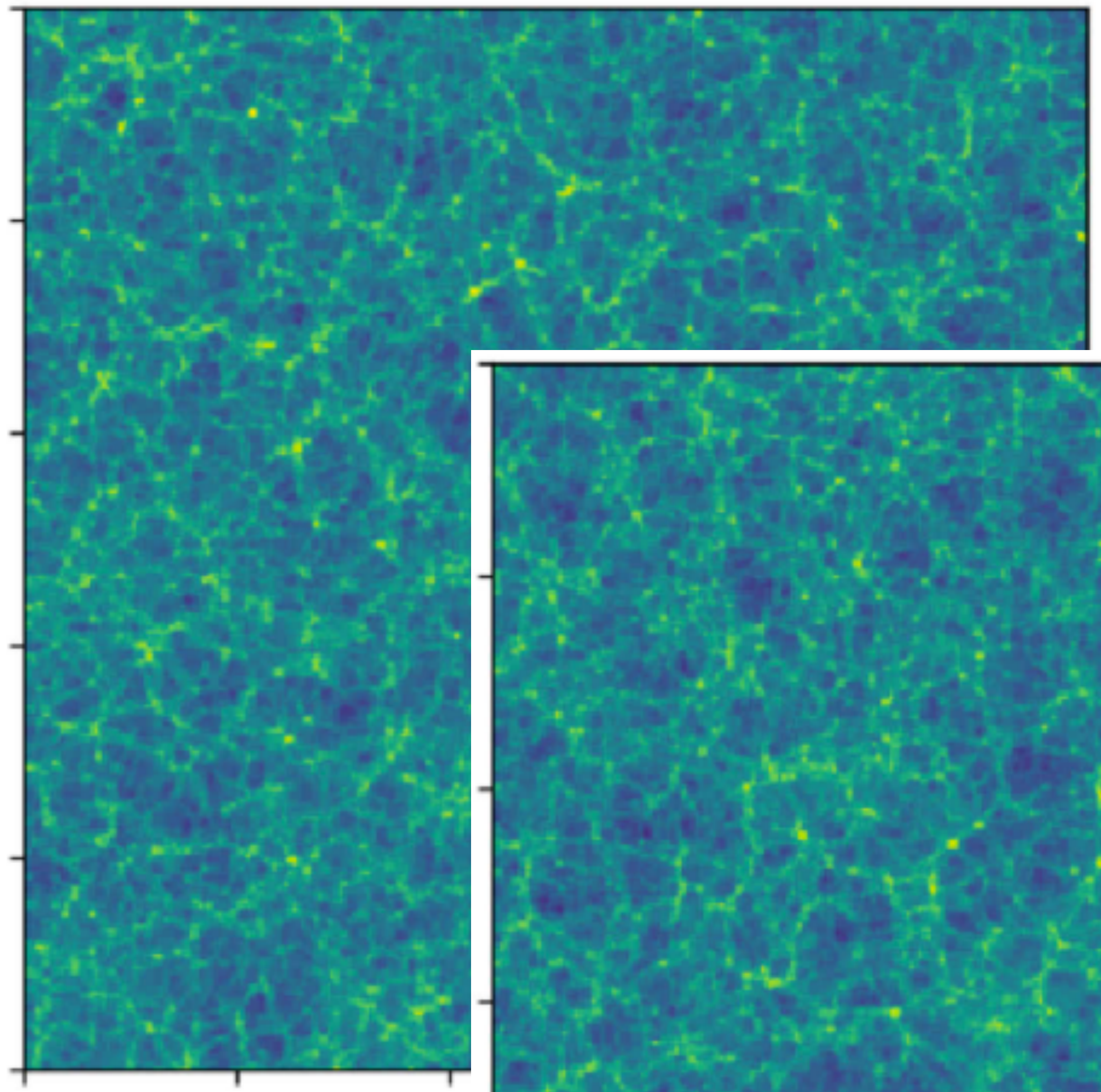
Cheng, Allys, Morel, et al. (in prep)



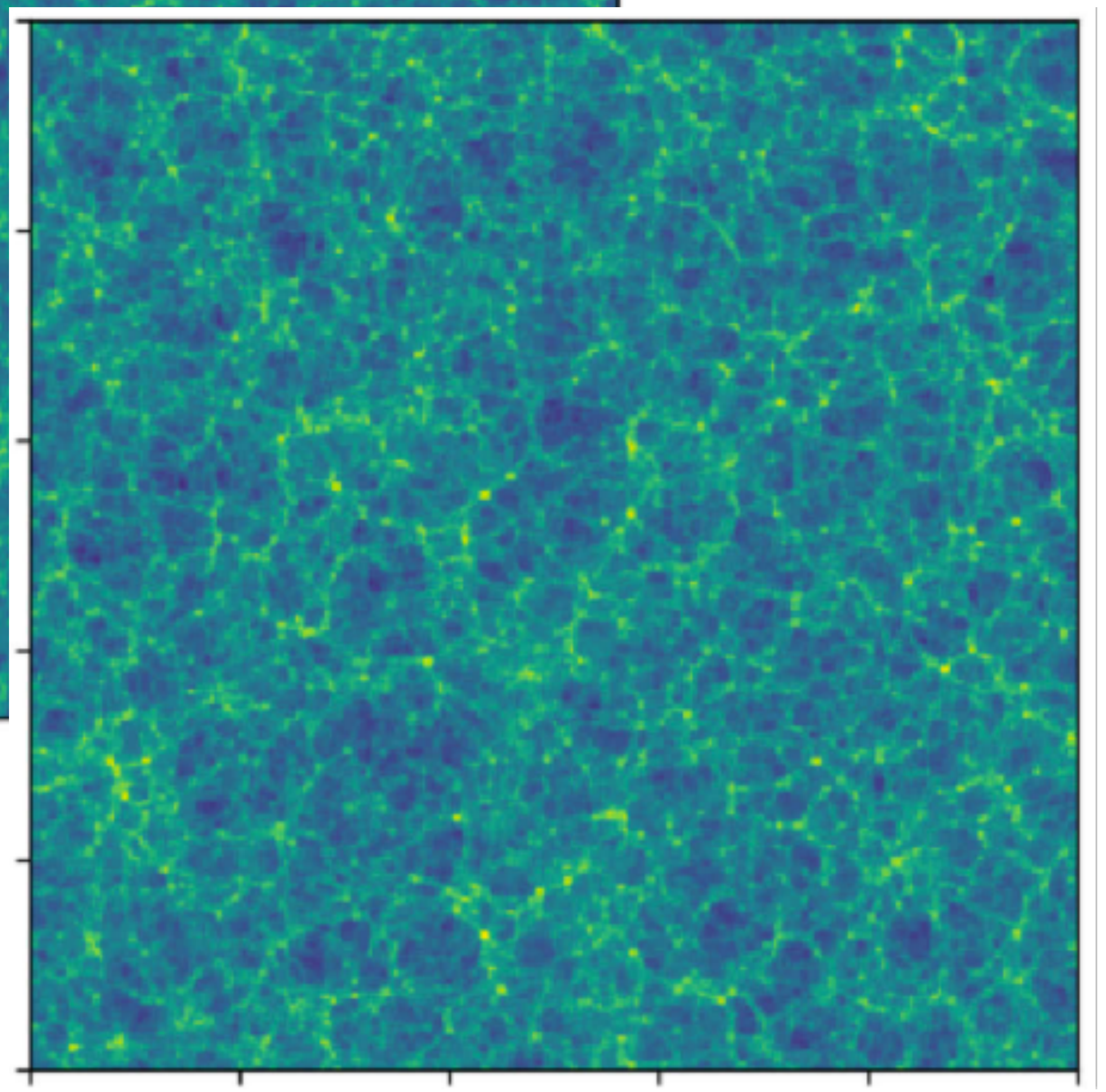


phase harmonic transform

original



synthesized



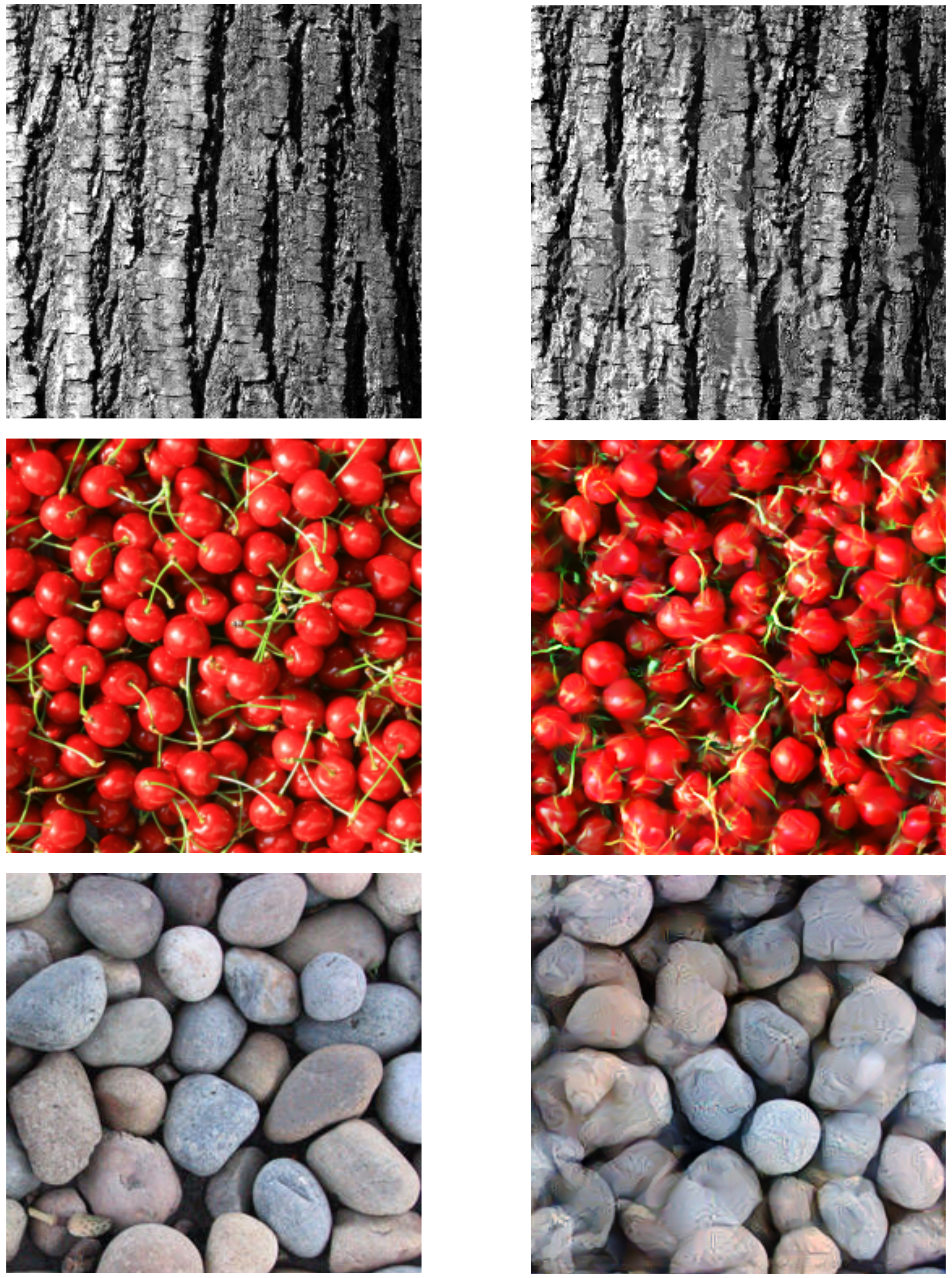
Allys et al (2020)

with ~6,000 coefficients
measured from 30 simulations

phase harmonic with spatial shifts

original synthesized

Brochard & Mallat
(submitted)



with 30k to 300k coefficients

In some cases, dimensionality reduction may provide us with a compact description

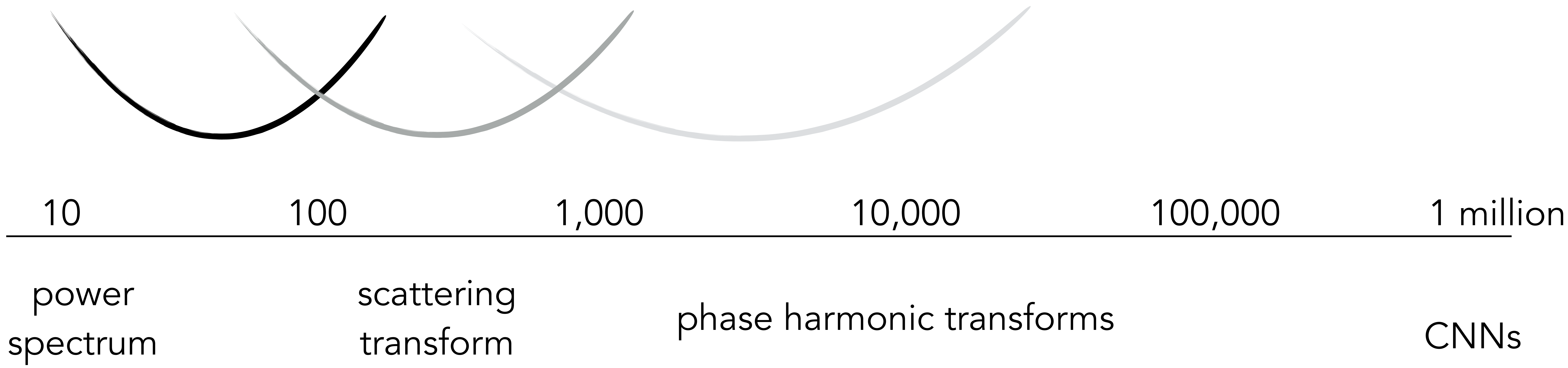
“mathematical” neural networks: how to use them?

For scientific data analysis, one wants to
- maximize expressivity
- minimize the number of parameters

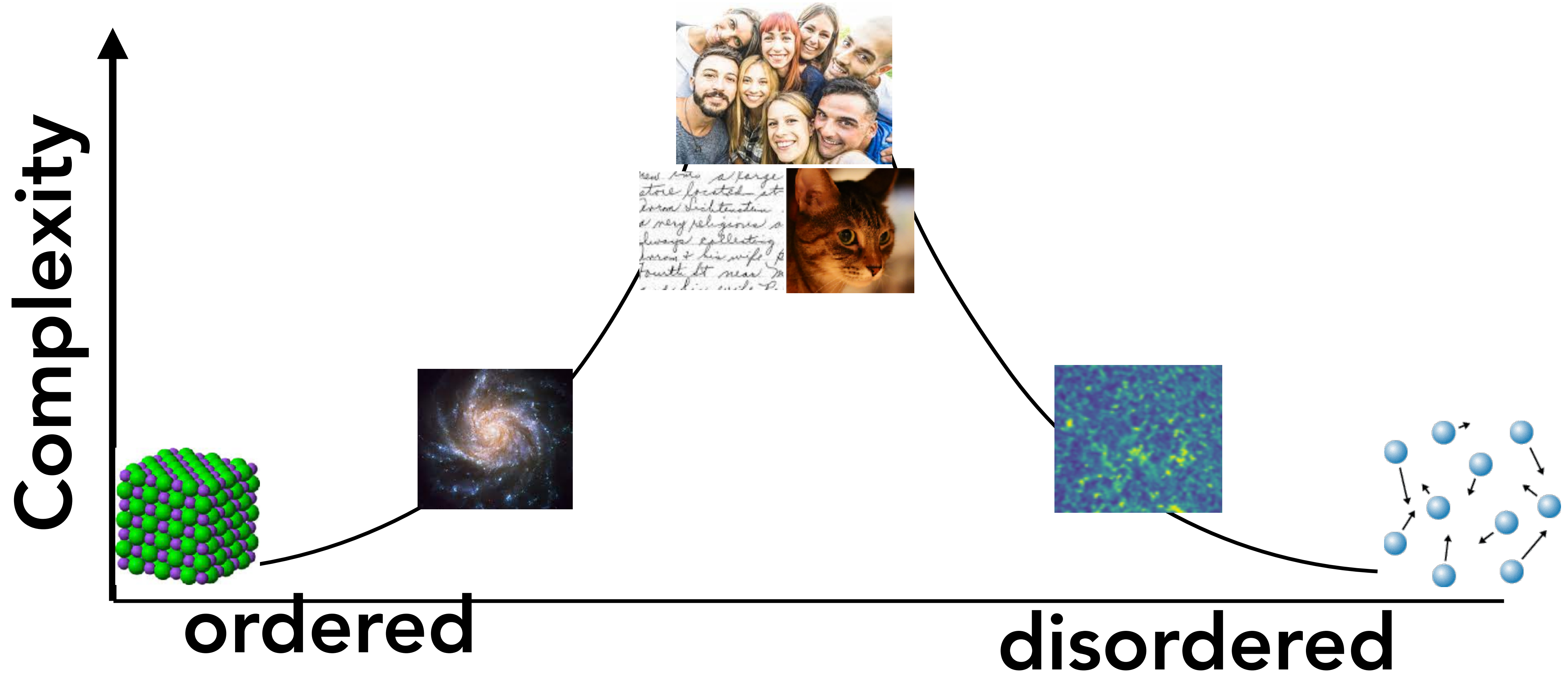
→ **there is a sweet spot somewhere**

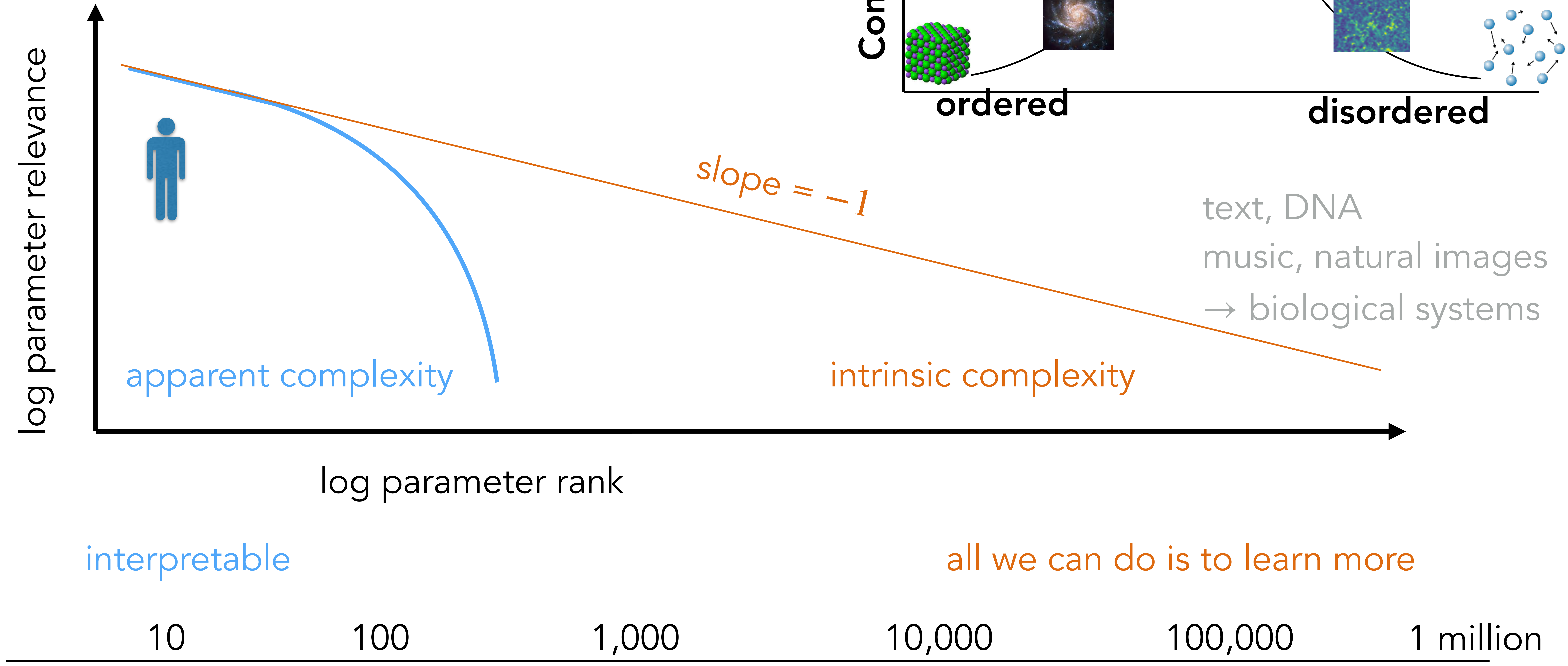
These transforms are data agnostic.
To be generic, they produce a lot of coefficients

→ **dimensionality reductions may compactify the description**



On interpretability and the limits of science





Key points

- We now have a range of statistical estimators for stationary fields

10	100	1,000	10,000	100,000	million
moment-based		mathematical networks			trained neural networks

- For scientific analyses, we want to
  - maximize expressivity
  - minimize the number of parameters

- A class of systems appear to exhibit unbounded intrinsic complexity.
  Their summary statistics/models can be arbitrarily large and beyond human scale.
  Do they still fall within the scope of Science?