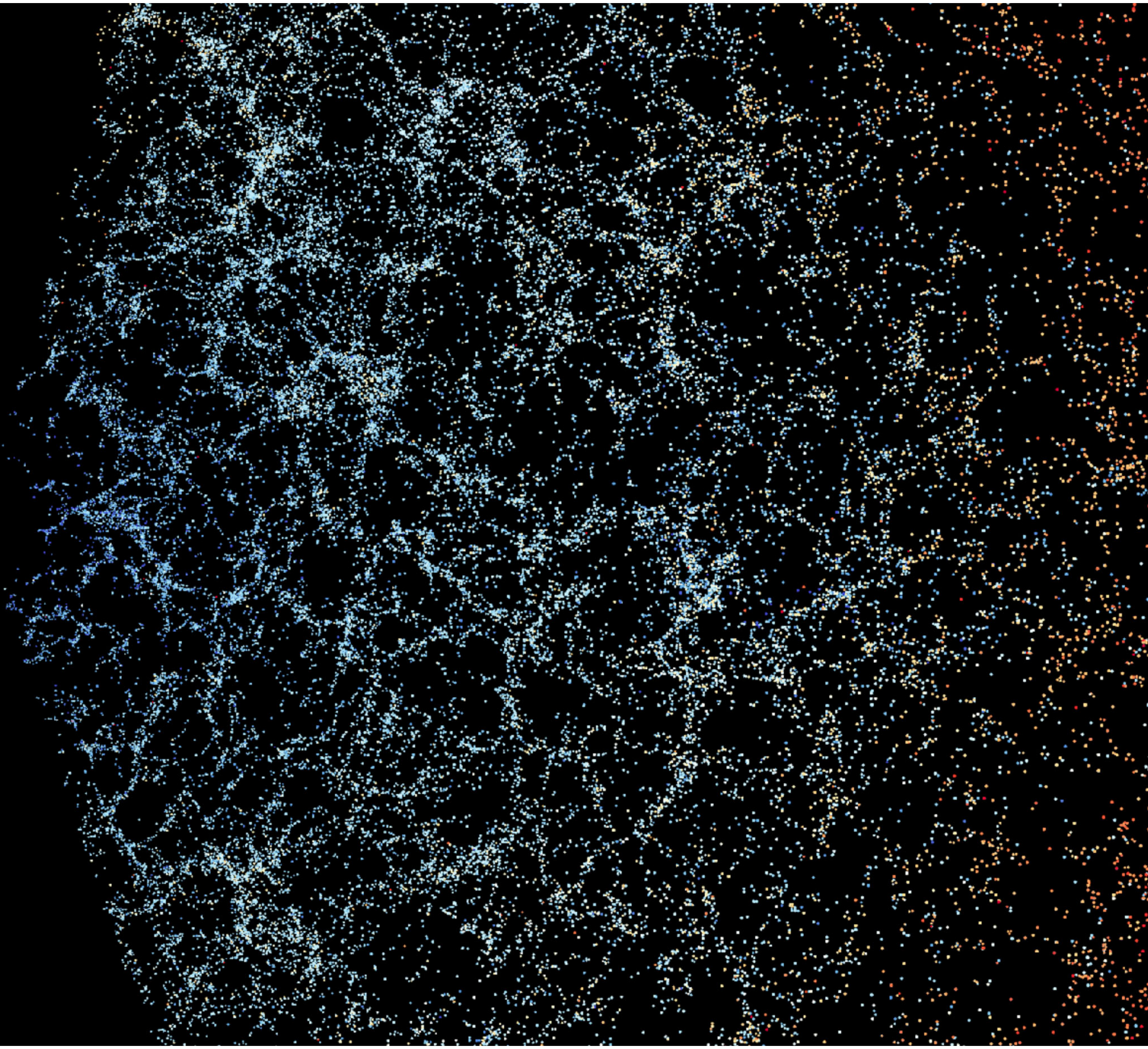
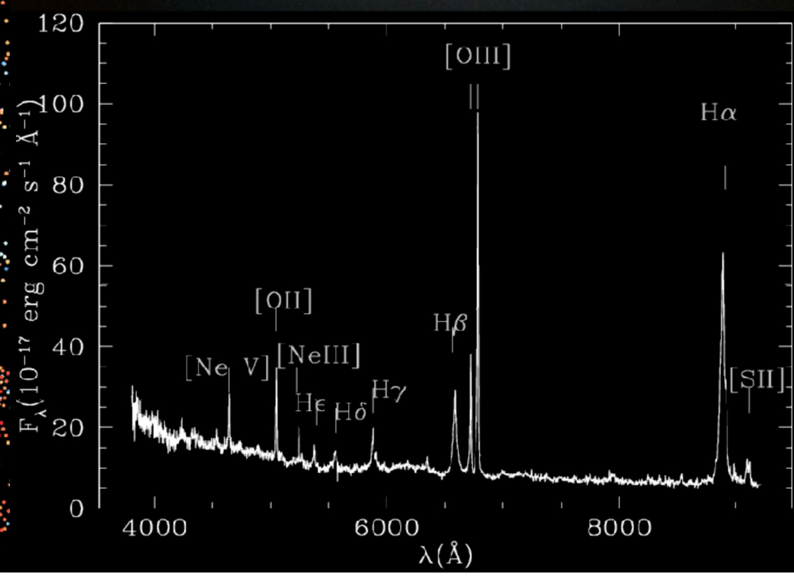
Statistics, machine learning, and astrophysics

Brice Ménard
Johns Hopkins University

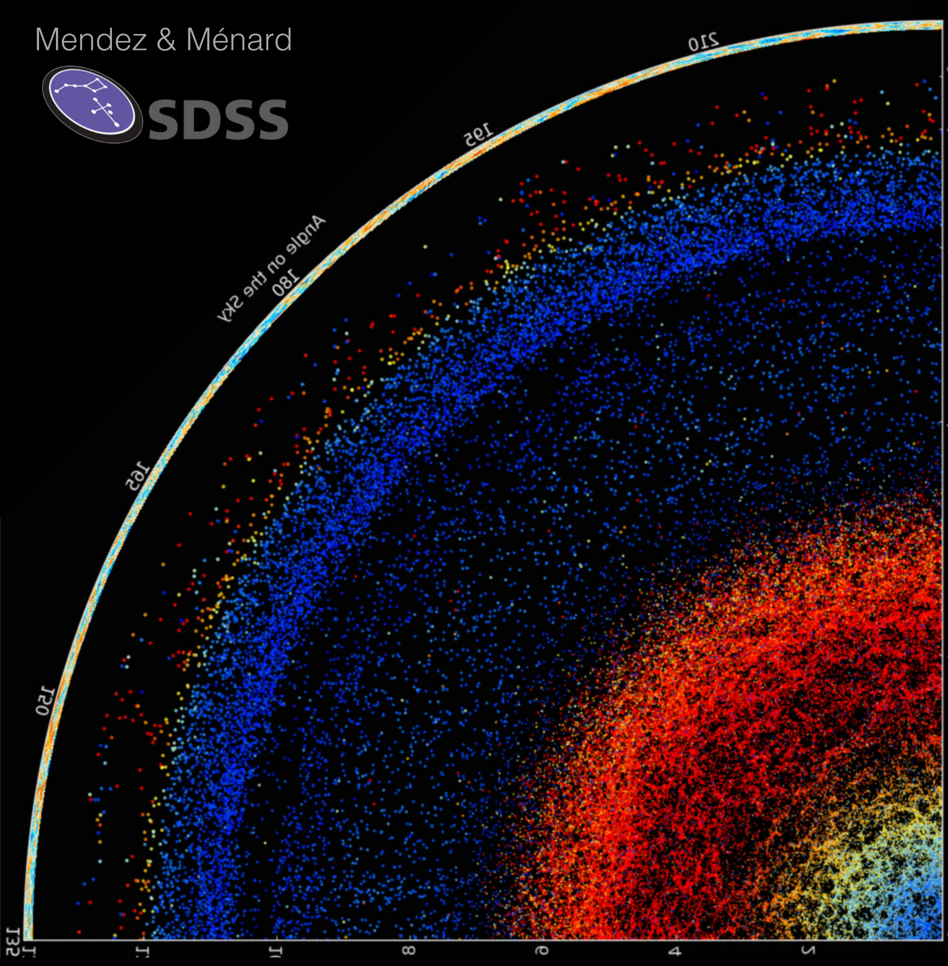
October 28 2019
NAS



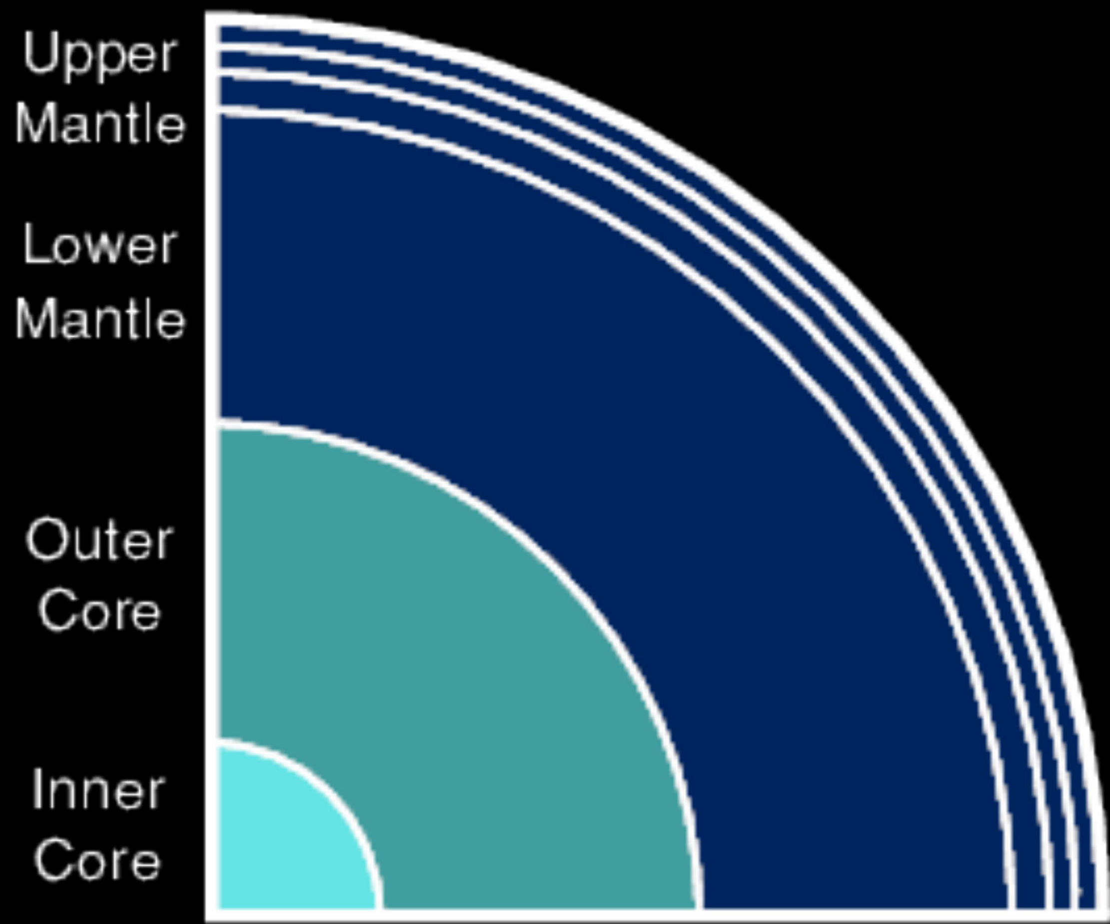


[OIII]
Hα
[OII]
Hβ
[Ne V]
[NeIII]
Hγ
Hε
Hδ
[SII]
F_λ(10^-17 erg cm^-2 s^-1 Å^-1)
0
20
40
60
80
100
120
4000
6000
8000
λ(Å)

Astrophysics



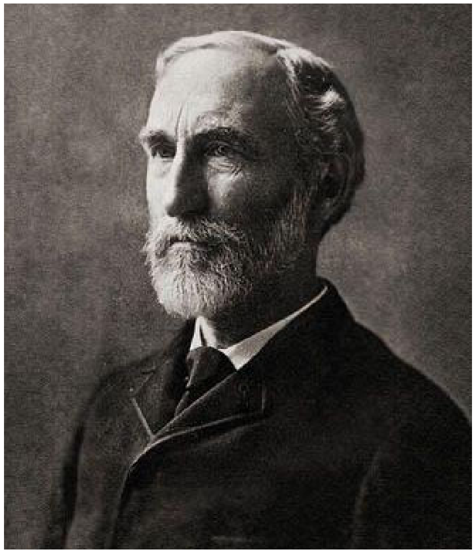
Geophysics



How do we do science?

"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)


Josiah Willard Gibbs (1839 – 1903)

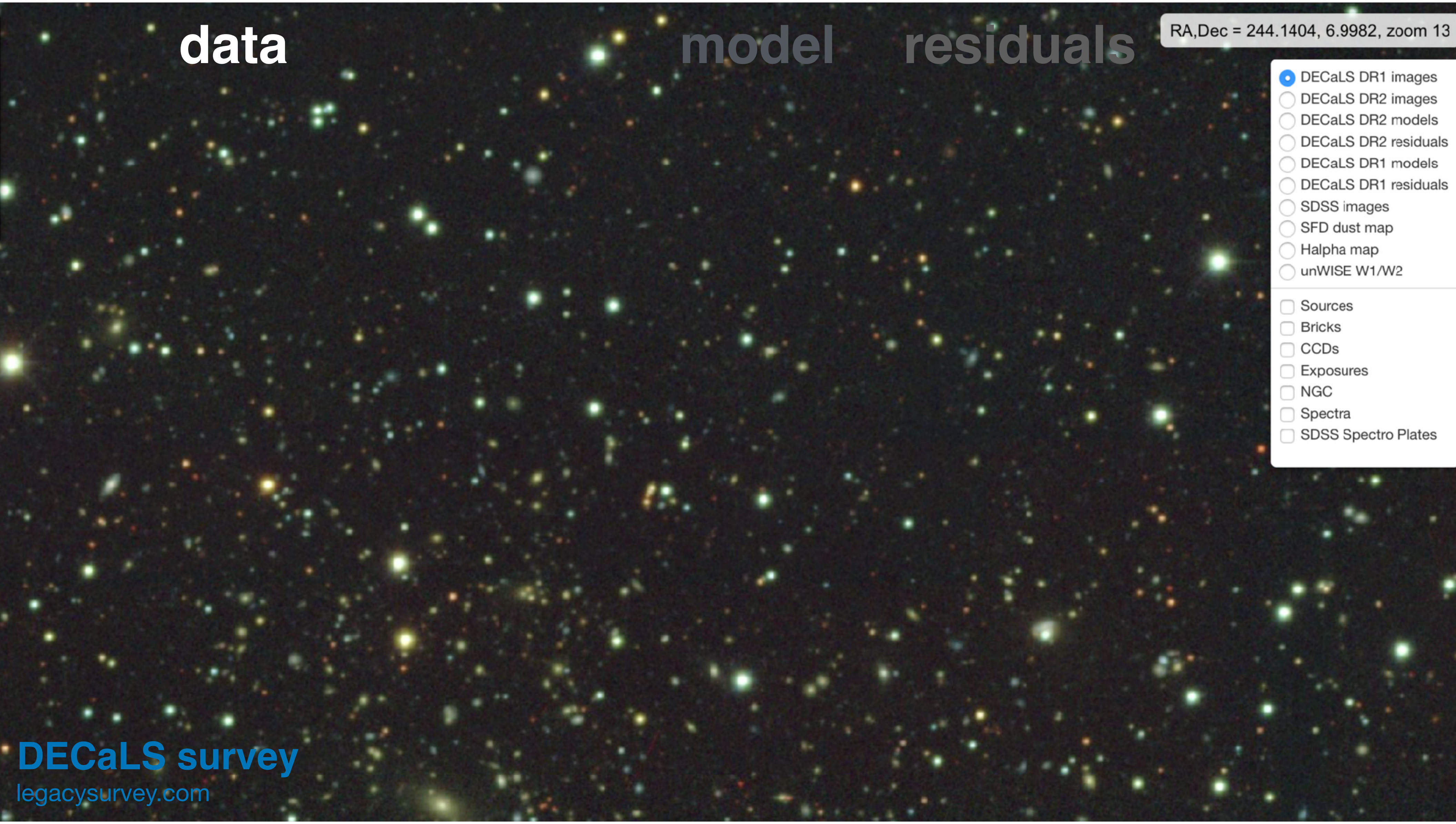
Complex world → Simplicity

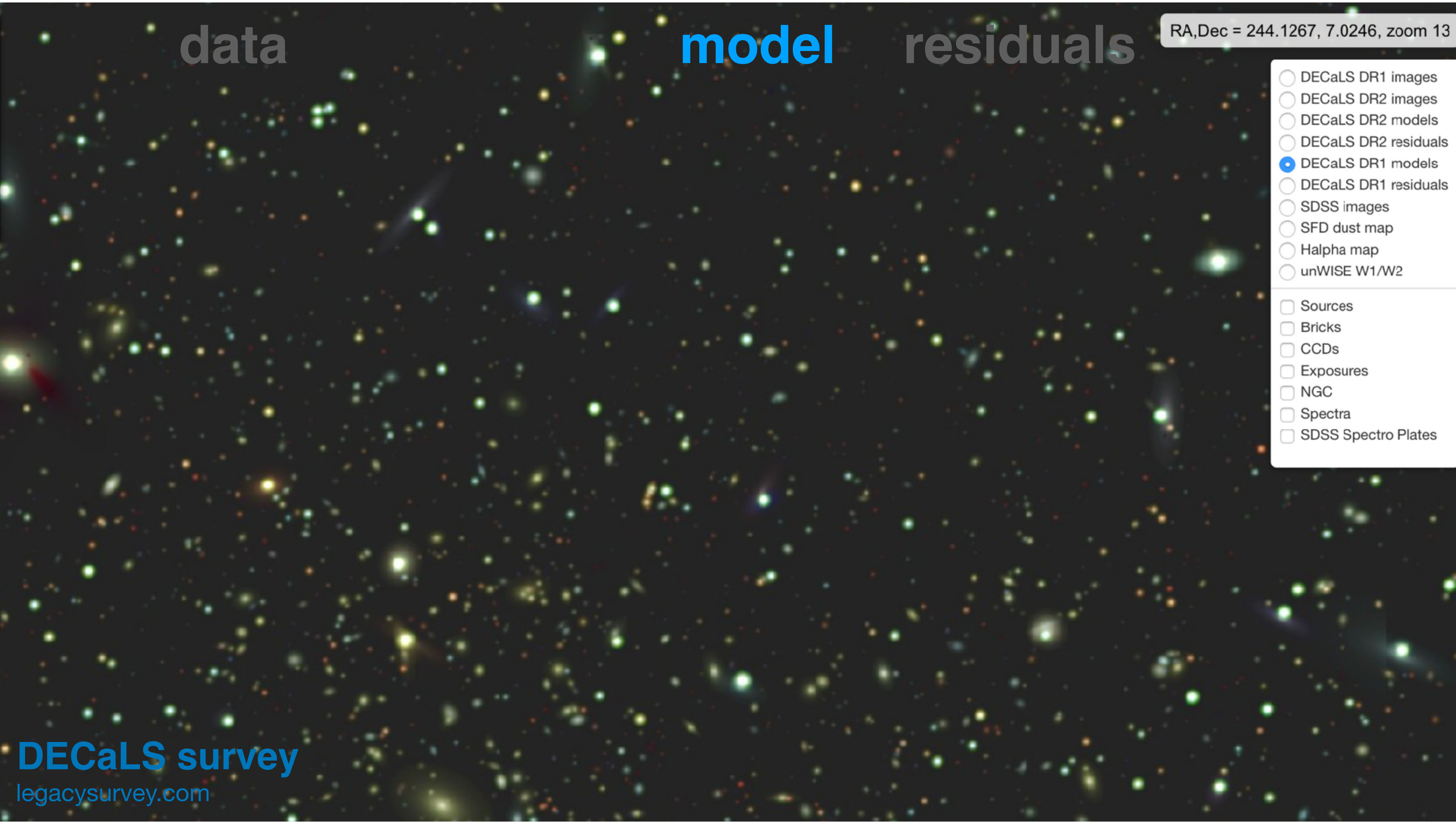
- Visual insight has been an important guide in science for centuries.

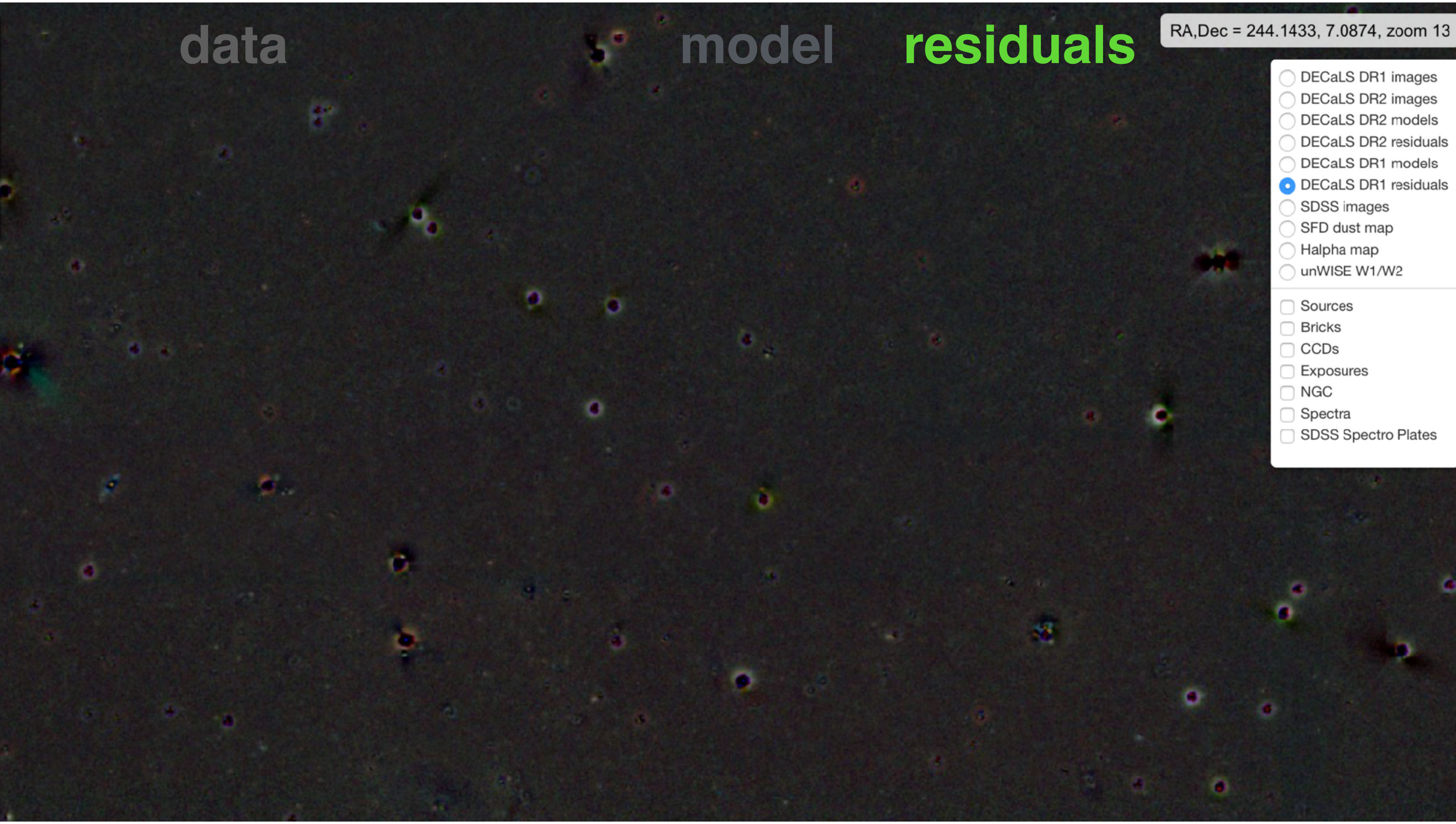
  → **visual insight is receding**

- We attempt to describe what we see using the language of physics & statistics.

  → **complexity can be a limitation**


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

model

object based

reduced dimensionality

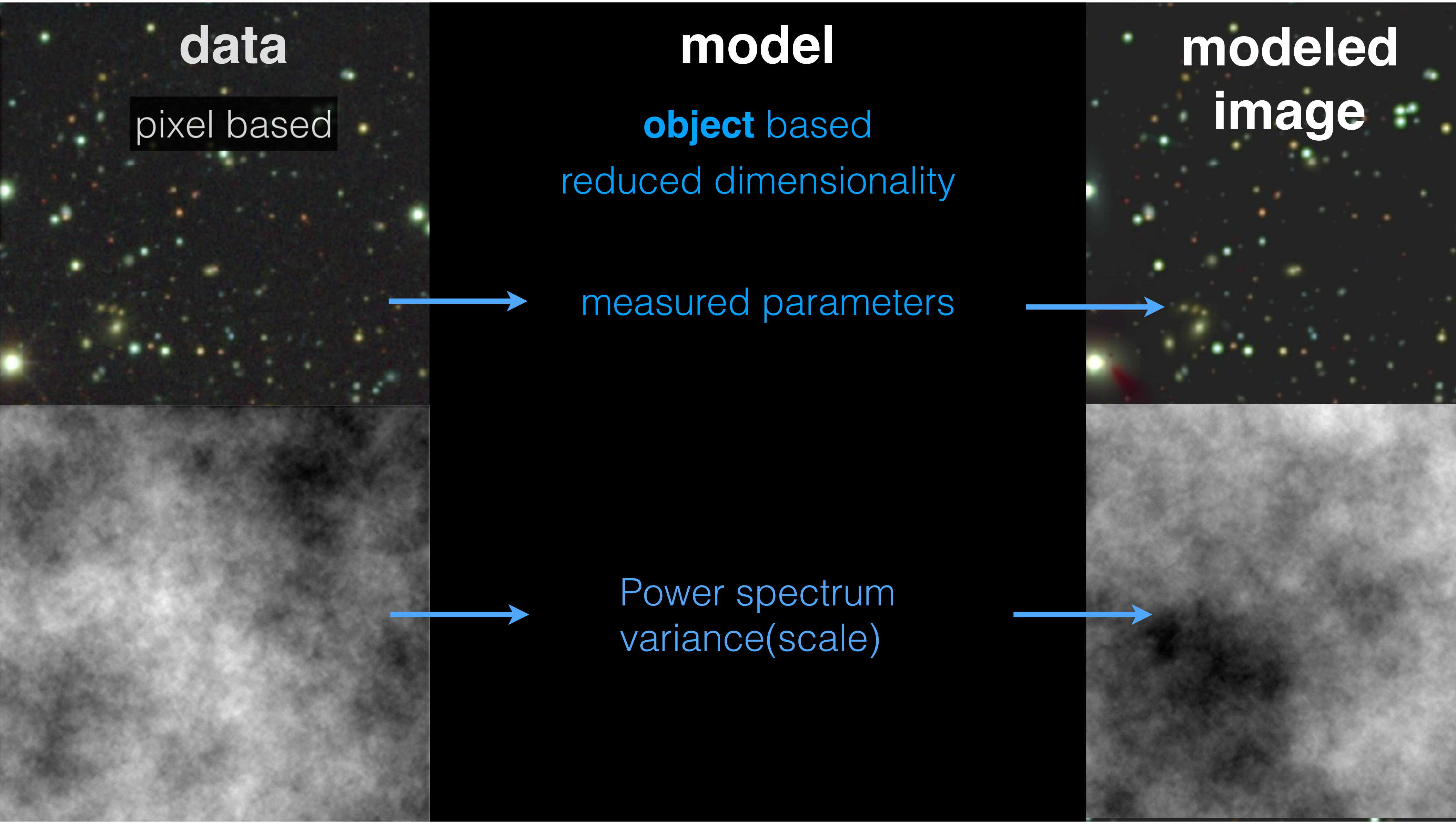
measured parameters

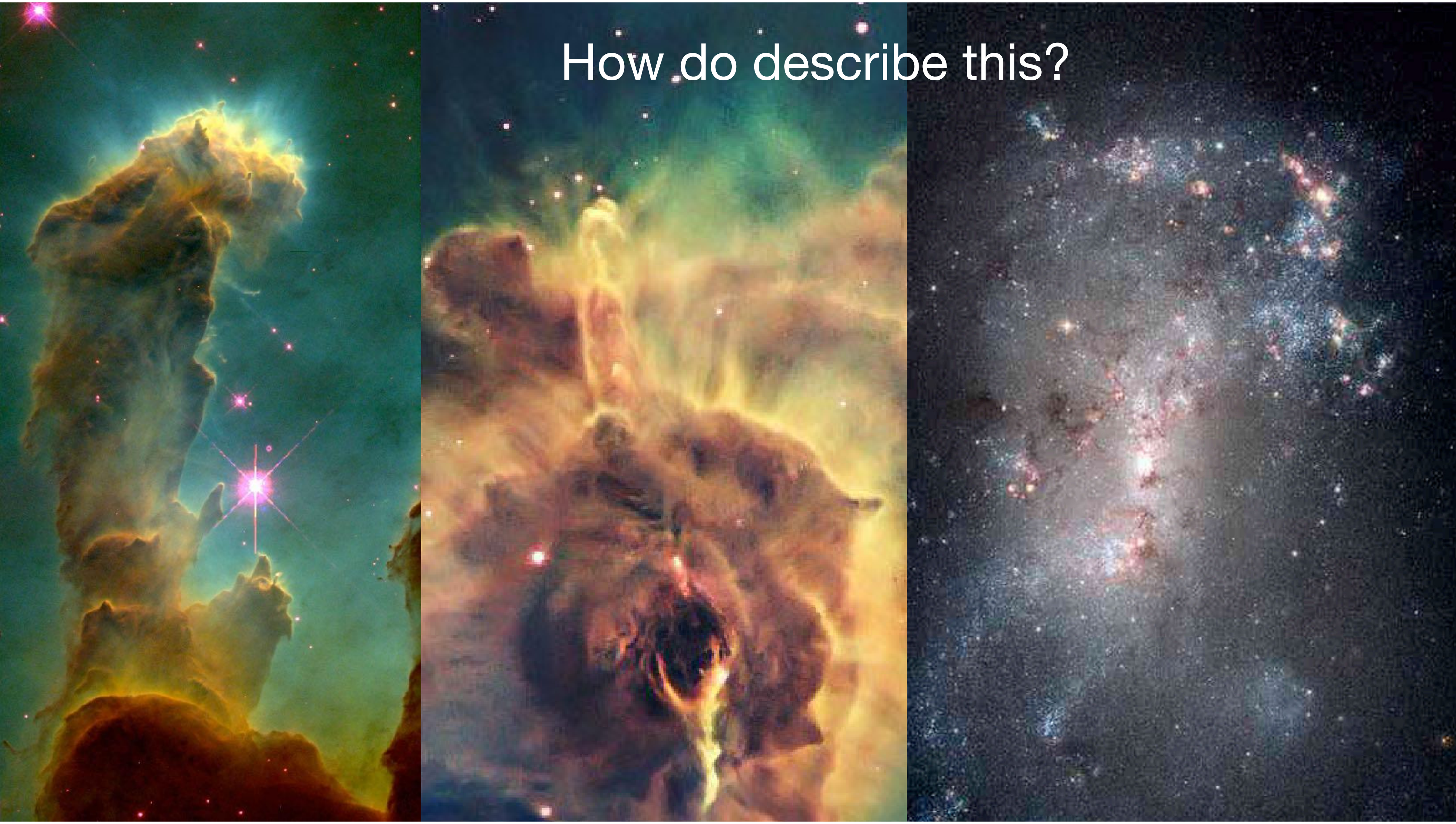
brightness
position
size
ellipticity
colors…

27.91	27.67	27.49	27.35
15.21	14.98	14.80	14.66
10.67	10.46	10.29	10.16
8.466	8.260	8.102	7.976
7.191	6.993	6.840	6.719
6.371	6.178	6.029	5.911
5.802	5.613	5.467	5.351
5.386	5.200	5.057	4.942

easier working environment

→ catalog/database (searchable)
→ correlations, constrain models

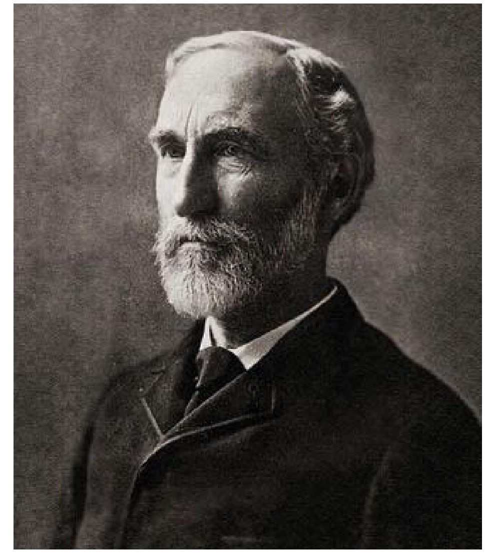

data
pixel based
model
object based
reduced dimensionality
measured parameters
Power spectrum
variance(scale)
modeled
image


How do describe this?

How do we do science?

"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)



Josiah Willard Gibbs (1839 – 1903)

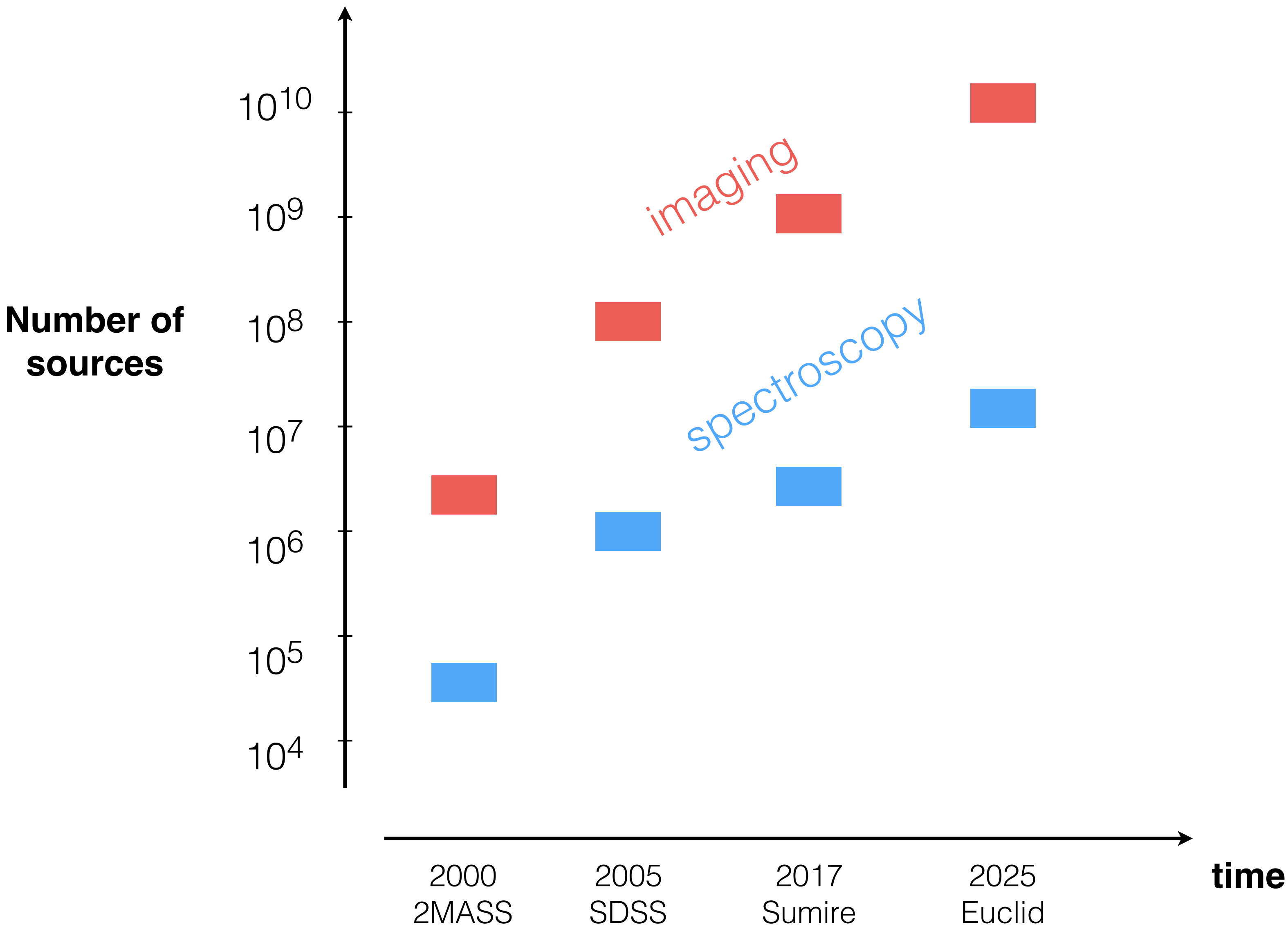
Complex world $\longrightarrow$ Simplicity

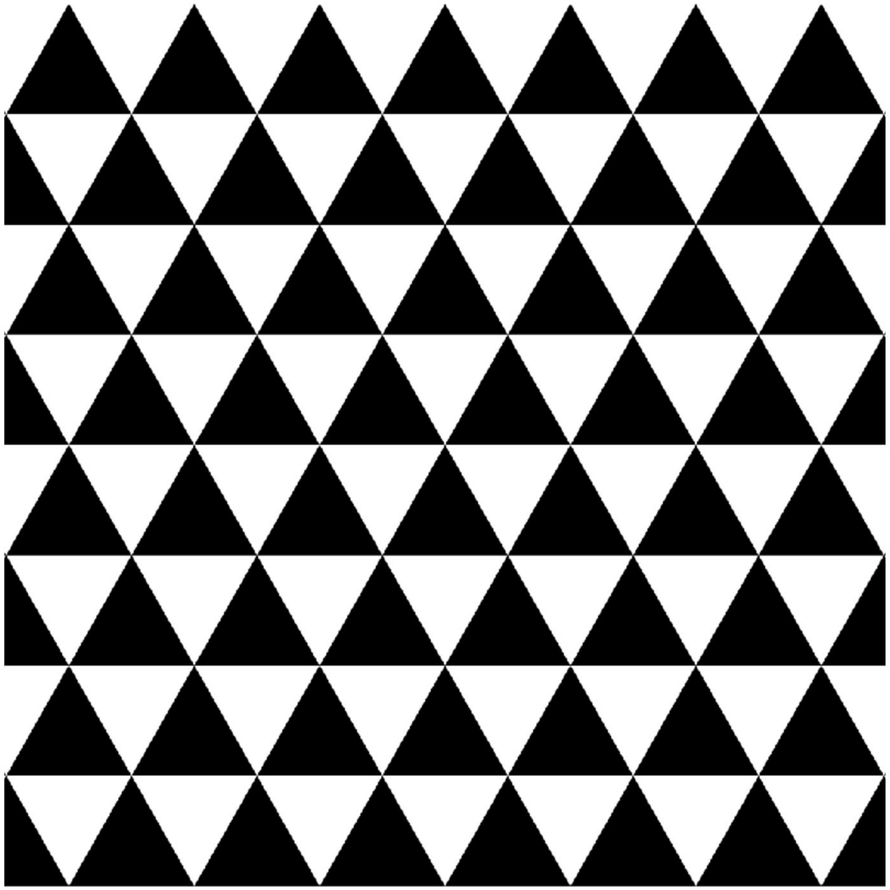
- Visual insight has been an important guide in science for centuries.

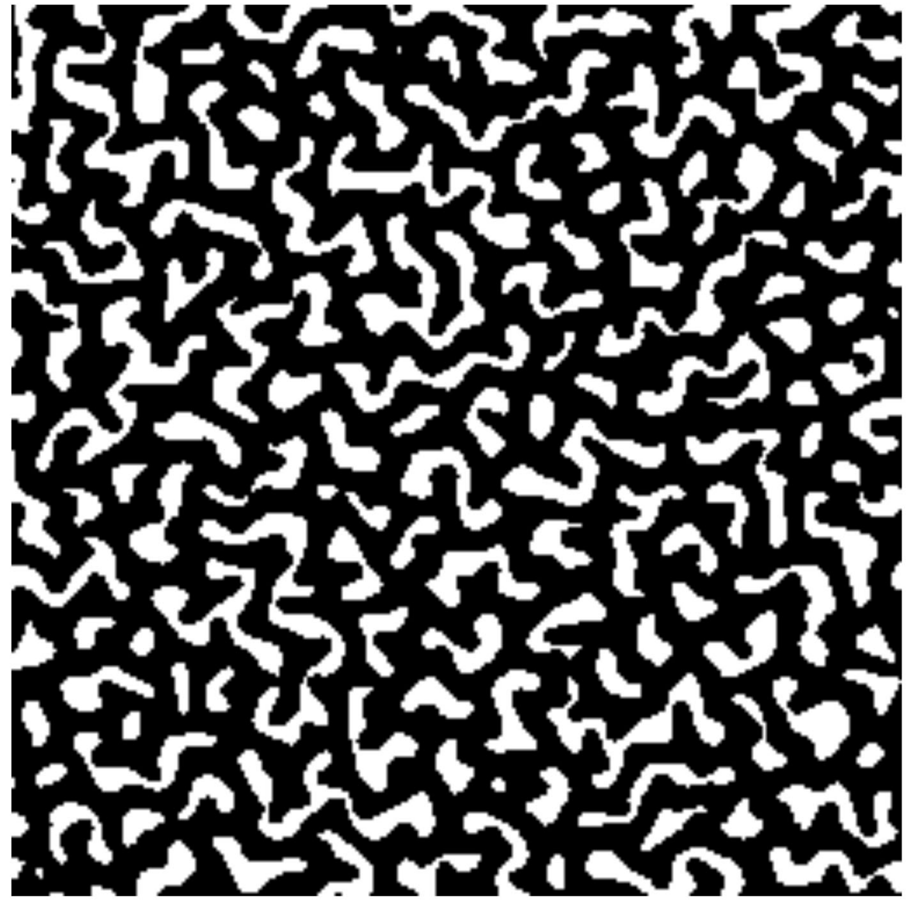
  → **visual insight is receding**

- We attempt to describe what we see using the language of physics & statistics.

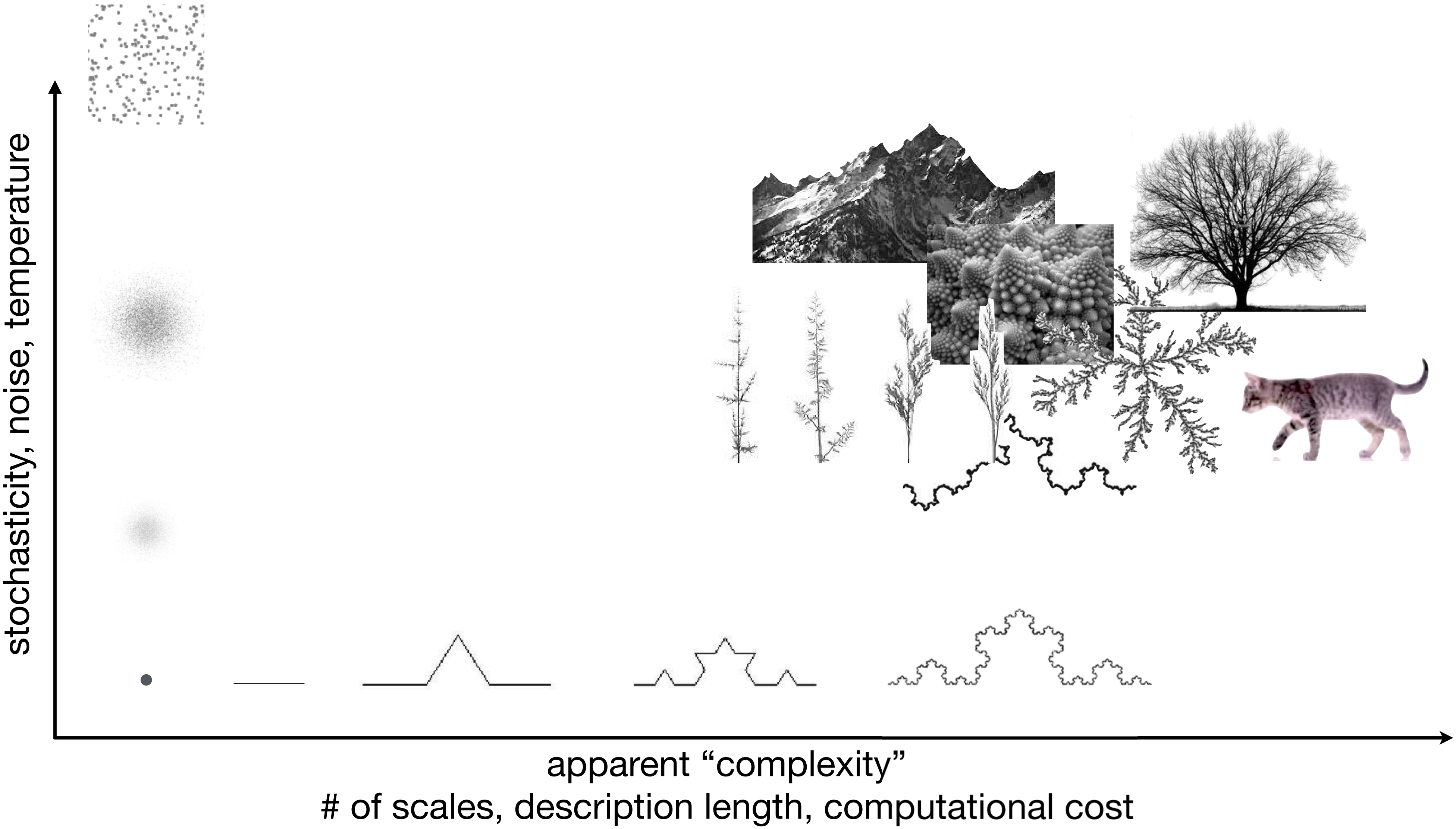
  → **complexity can be a limitation**

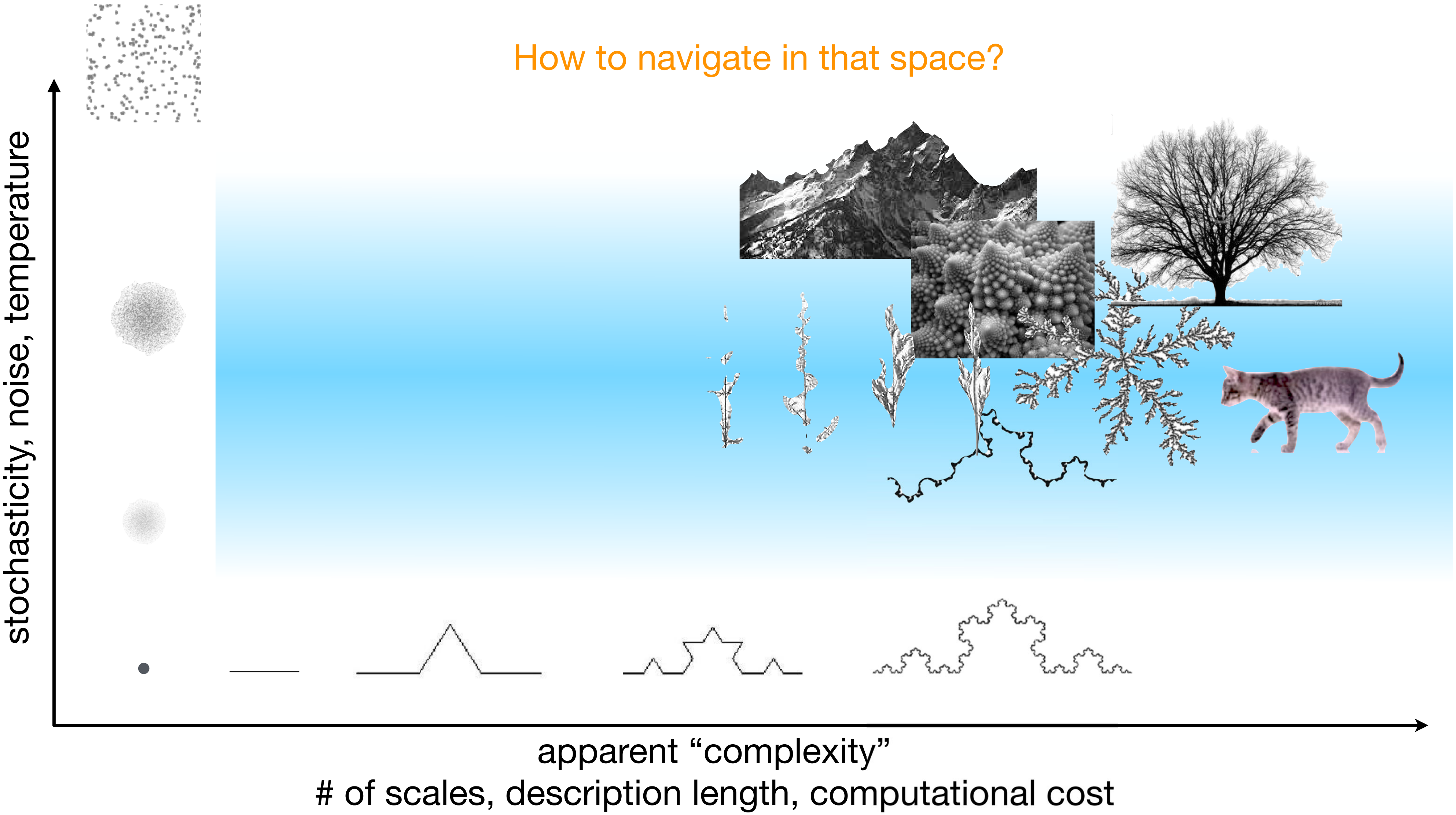

Number of sources
10^{10}
10^9
10^8
10^7
10^6
10^5
10^4
imaging
spectroscopy
2000
2MASS
2005
SDSS
2017
Sumire
2025
Euclid
time



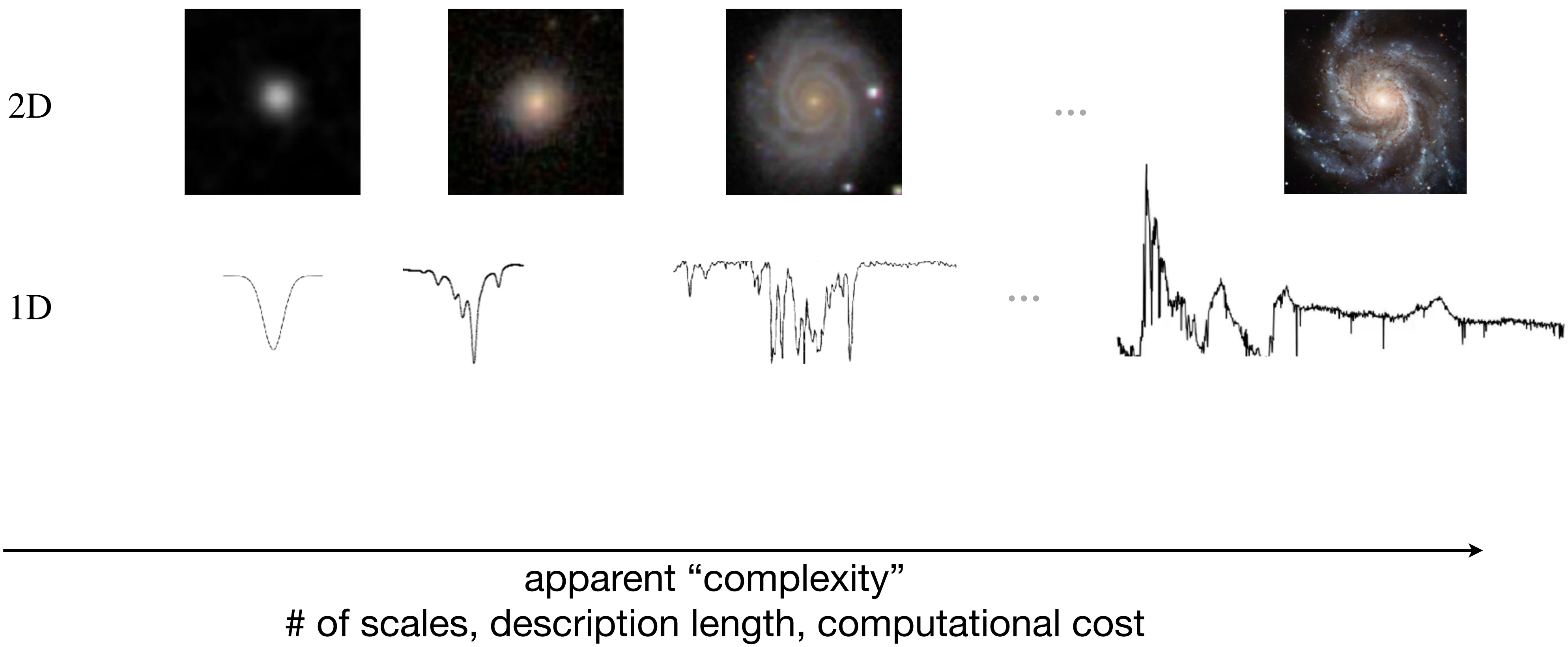


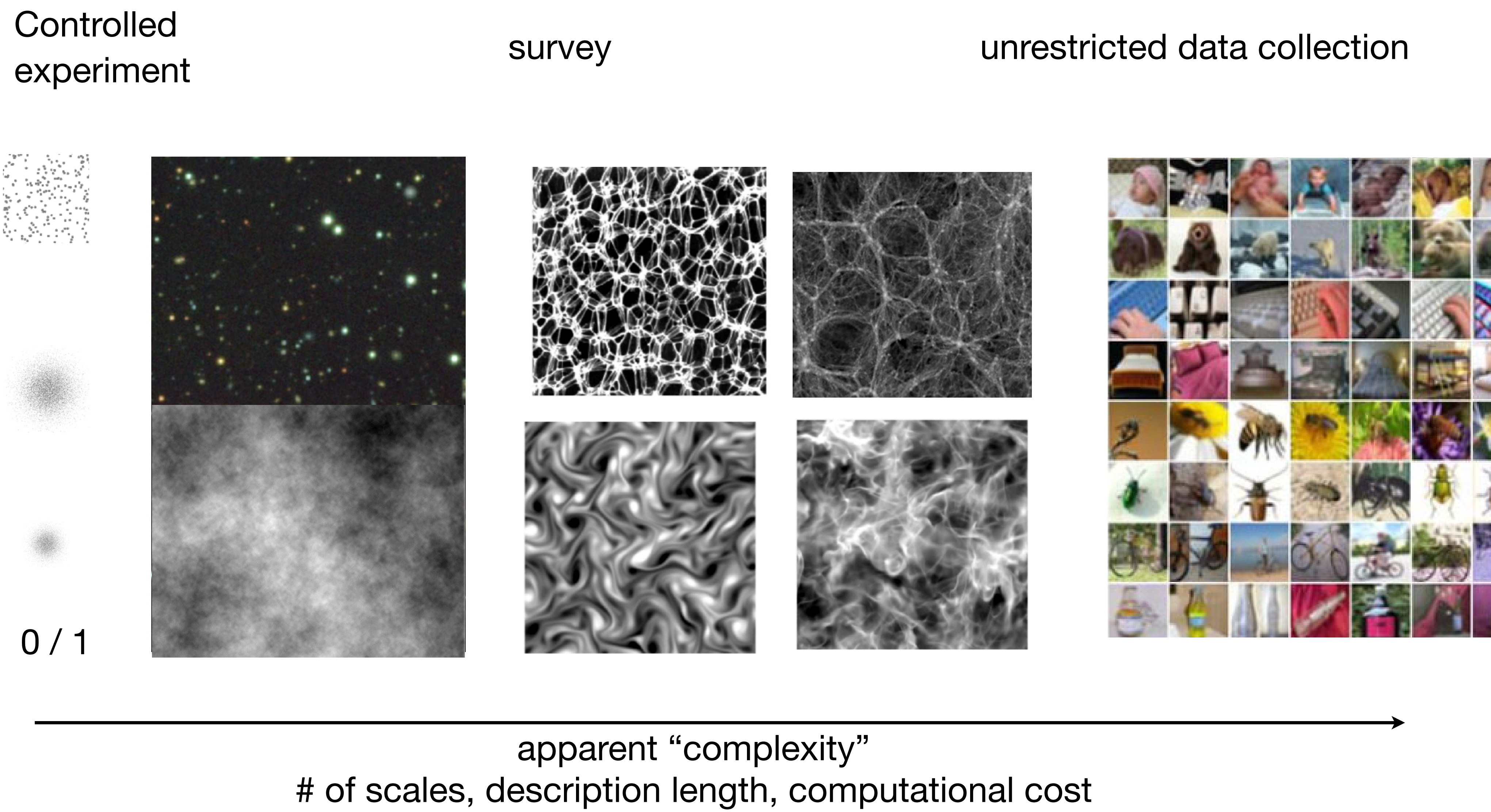
Language can be a serious limitation


stochasticity, noise, temperature
apparent “complexity”
# of scales, description length, computational cost


How to navigate in that space?
stochasticity, noise, temperature
apparent “complexity”
# of scales, description length, computational cost

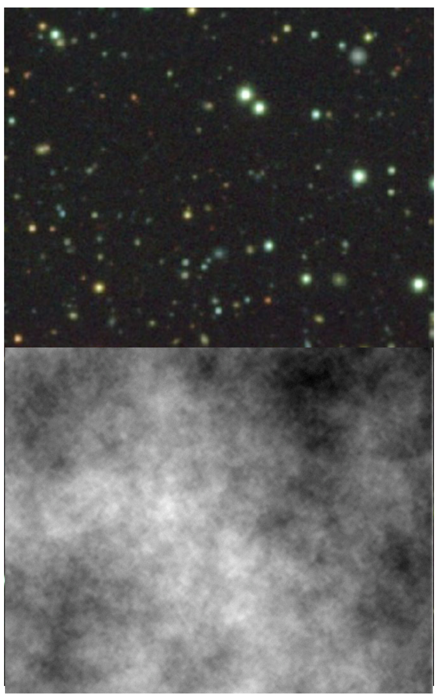
How to navigate in that space?

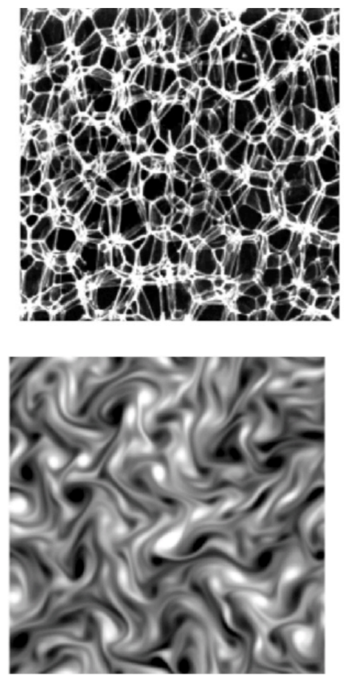


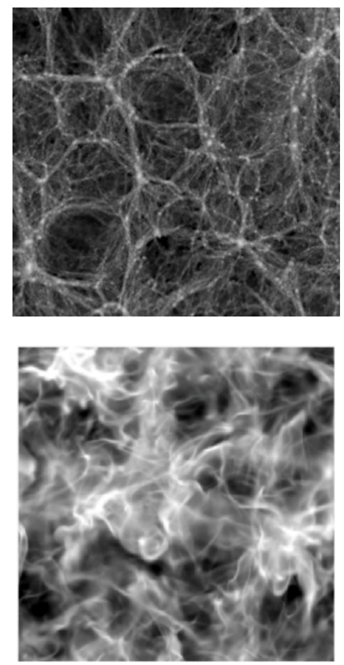

Controlled experiment
survey
unrestricted data collection
0 / 1
apparent “complexity”
# of scales, description length, computational cost

interpretability challenge, control of systematics

Gaussian statistics | manifold learning | dictionary learning | (convolutional) neural nets





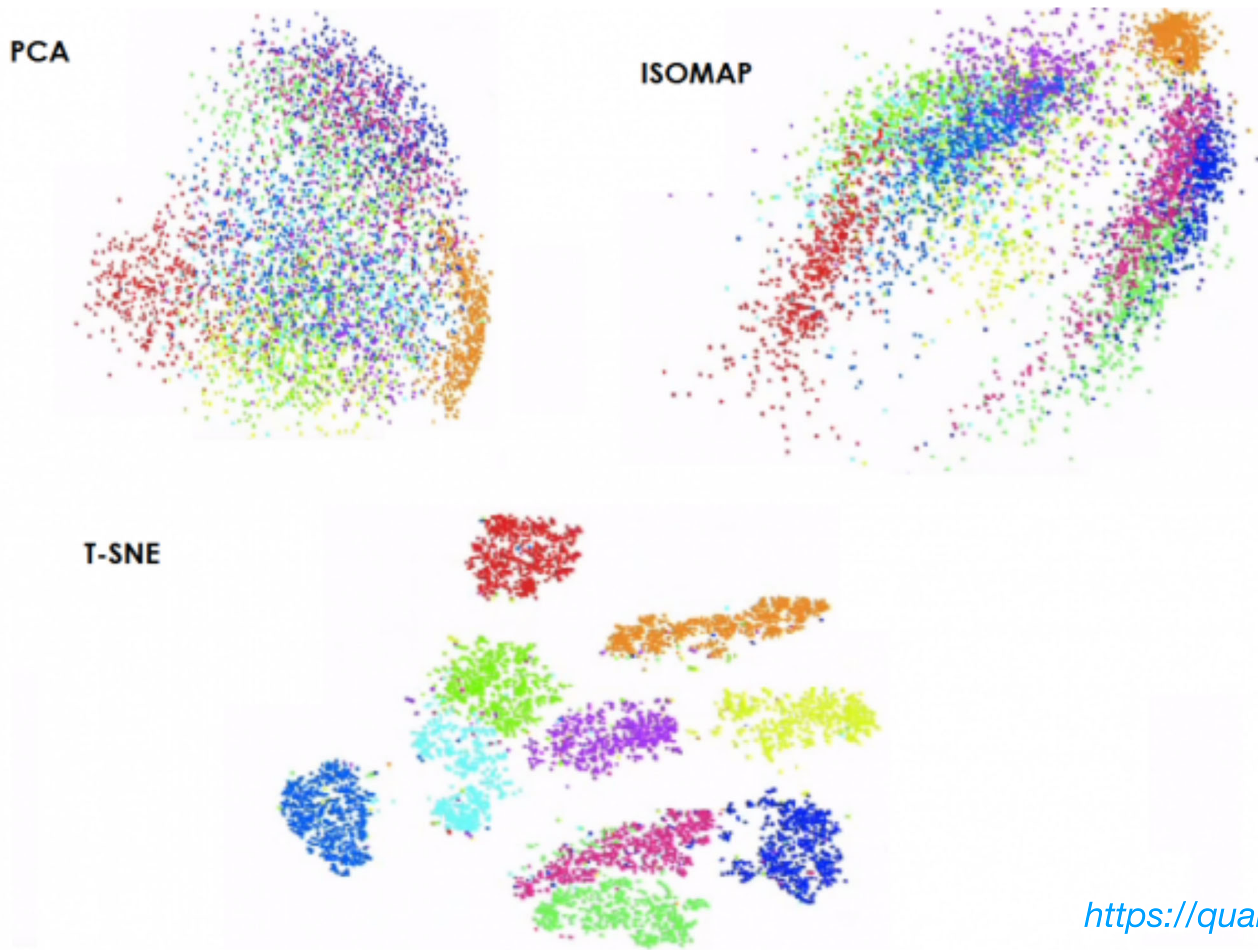


apparent “complexity”

# of scales, description length, computational cost

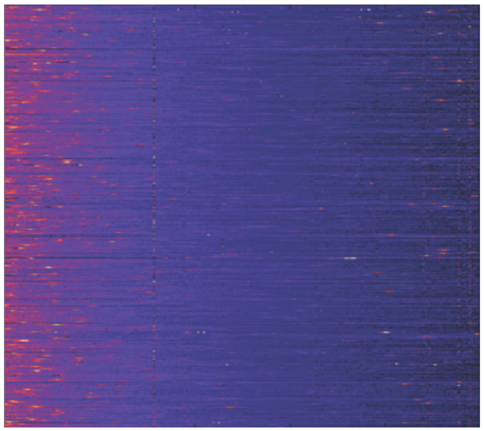
Gaussian statistics — manifold learning — dictionary learning — (convolutional) neural nets

complexity

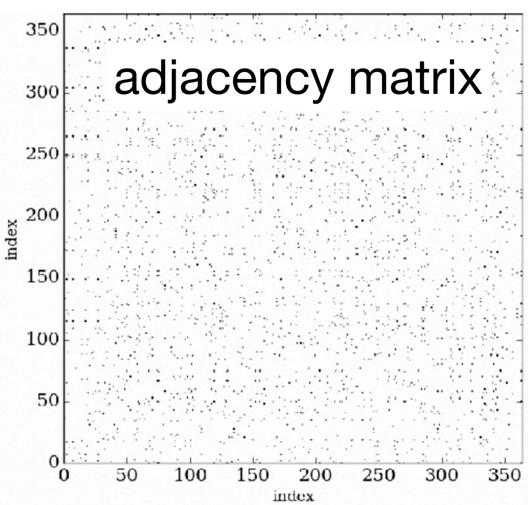


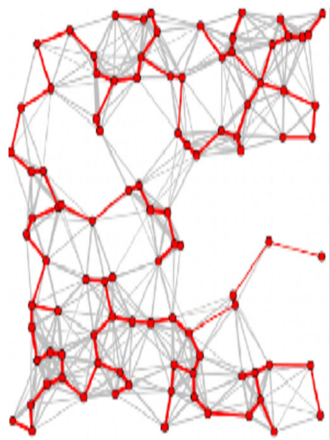
https://quantdare.com/tsne/

Gaussian statistics — **manifold learning** — dictionary learning — (convolutional) neural nets



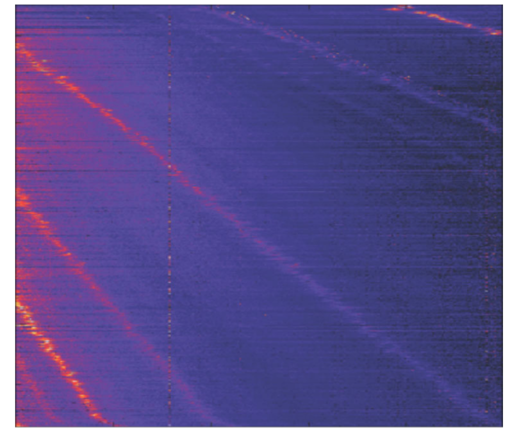
TheSequencer.org

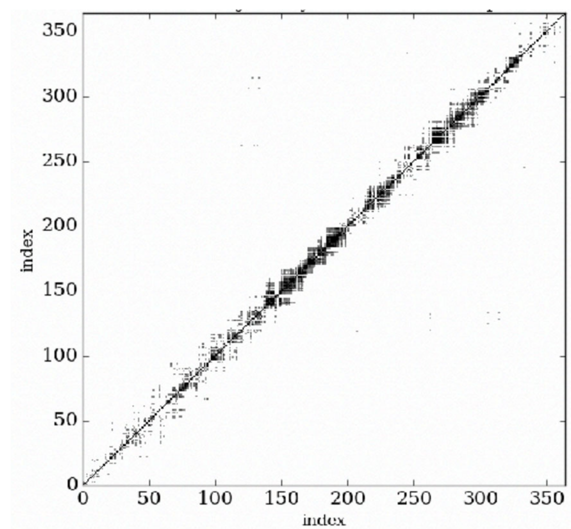




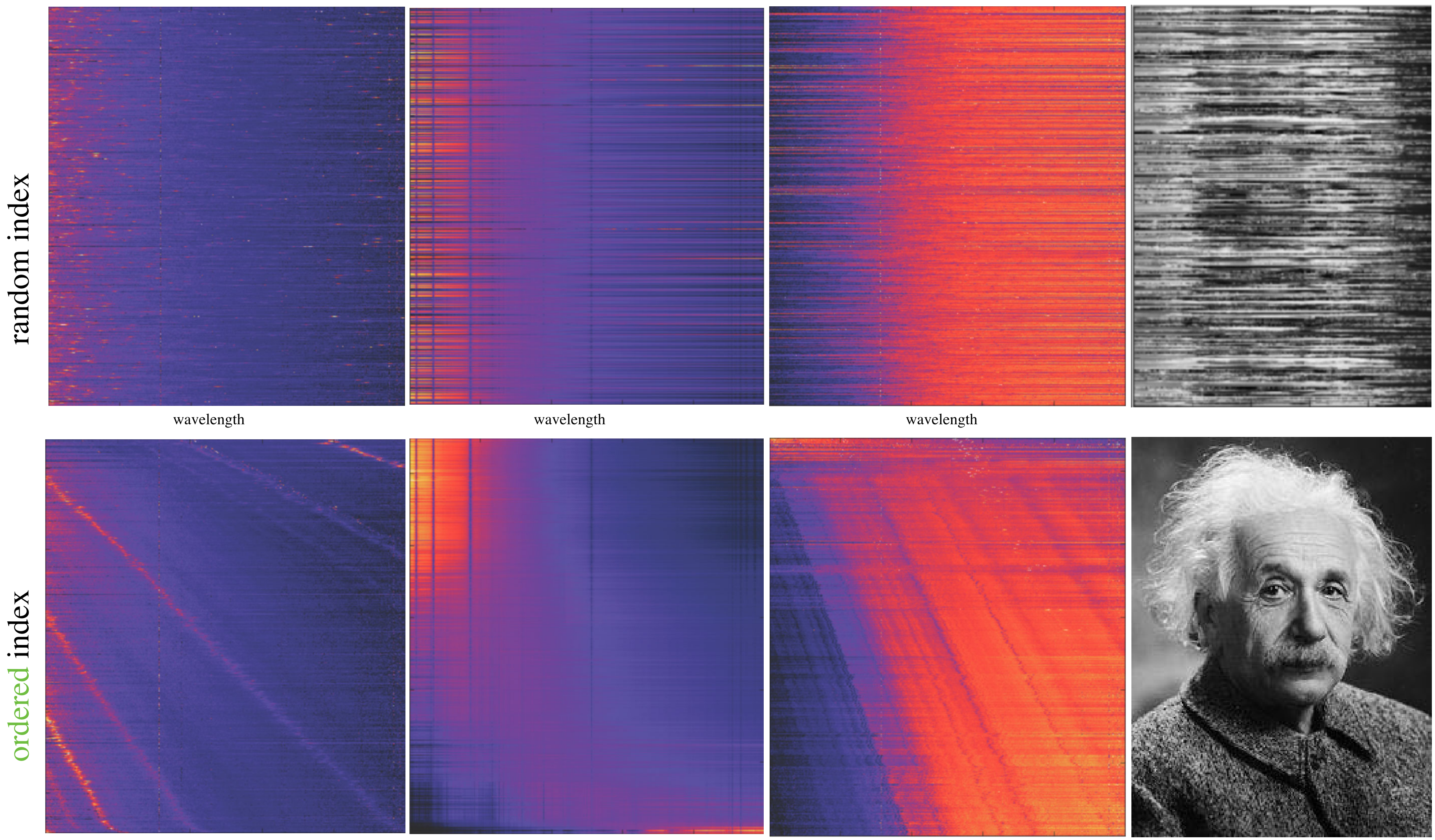
sequence found
using geometric properties

D. Baron & BM (in prep)

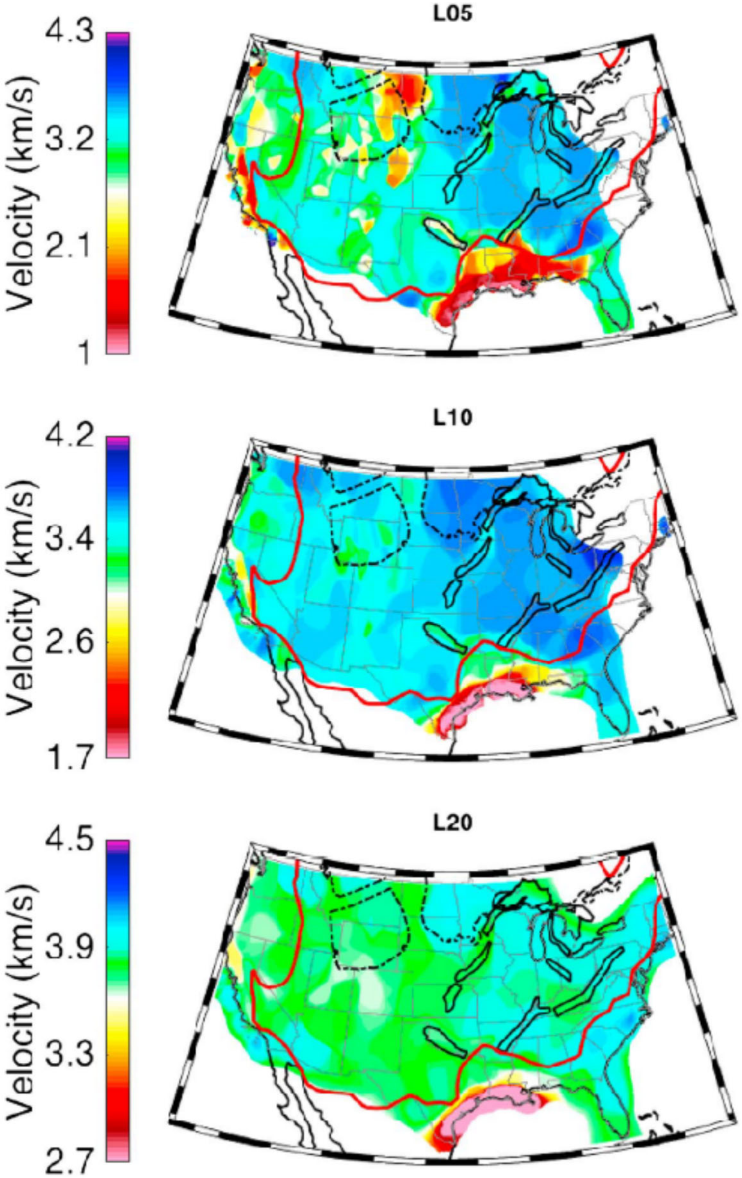


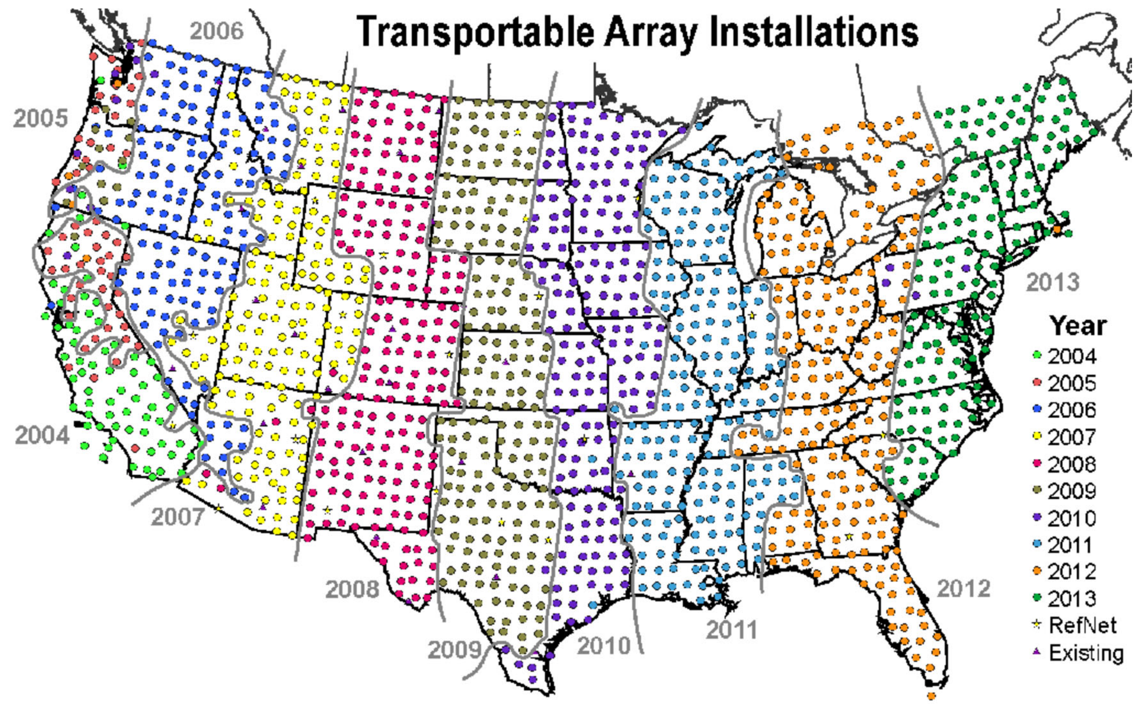


TheSequencer.org



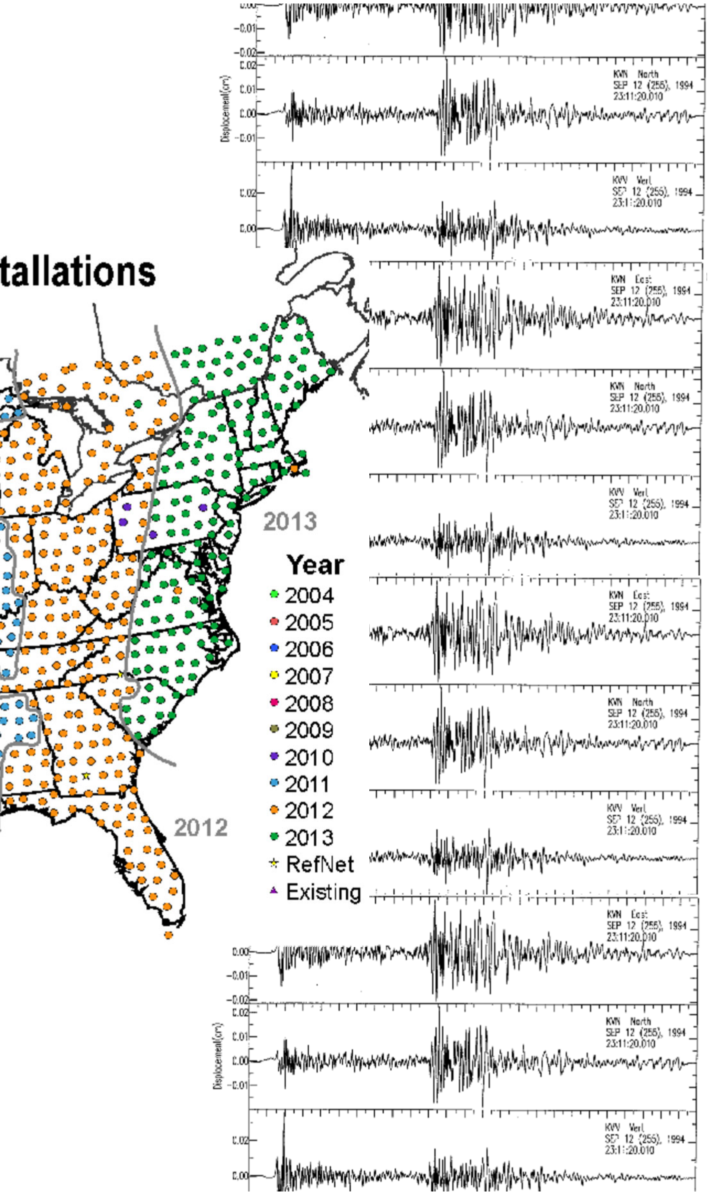
Map making in seismology



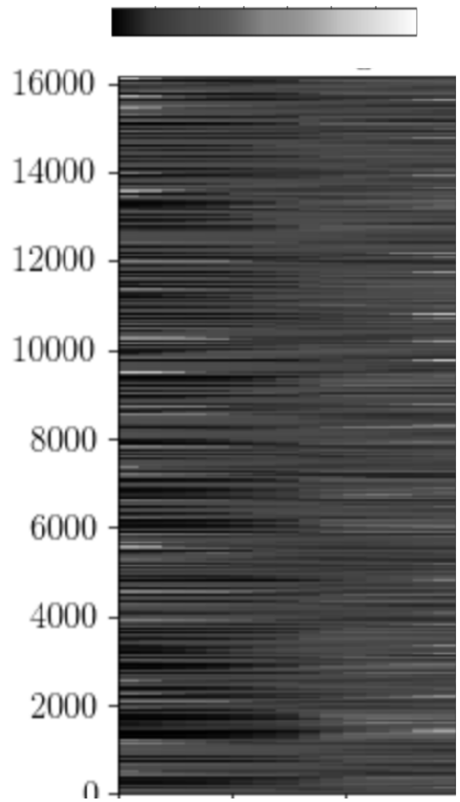


with V. Lekic (Univ. of Maryland)

Map making in seismology

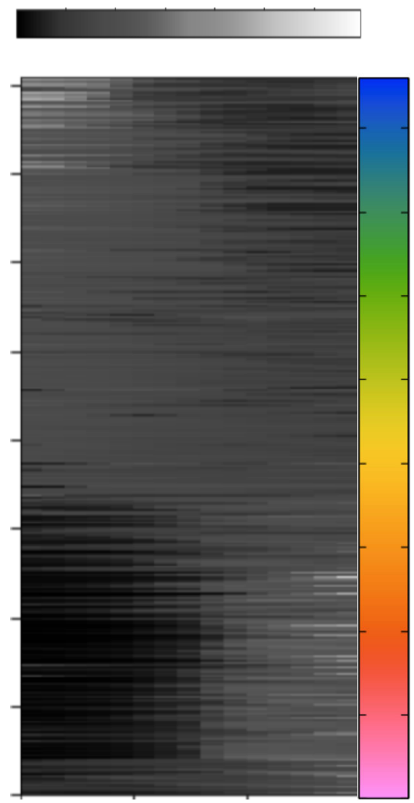


surface wave velocity



frequency

surface wave velocity
ordered by theSequencer.org

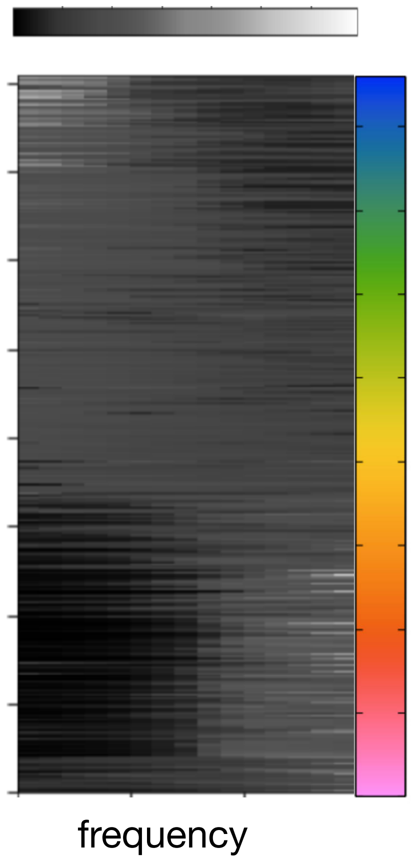


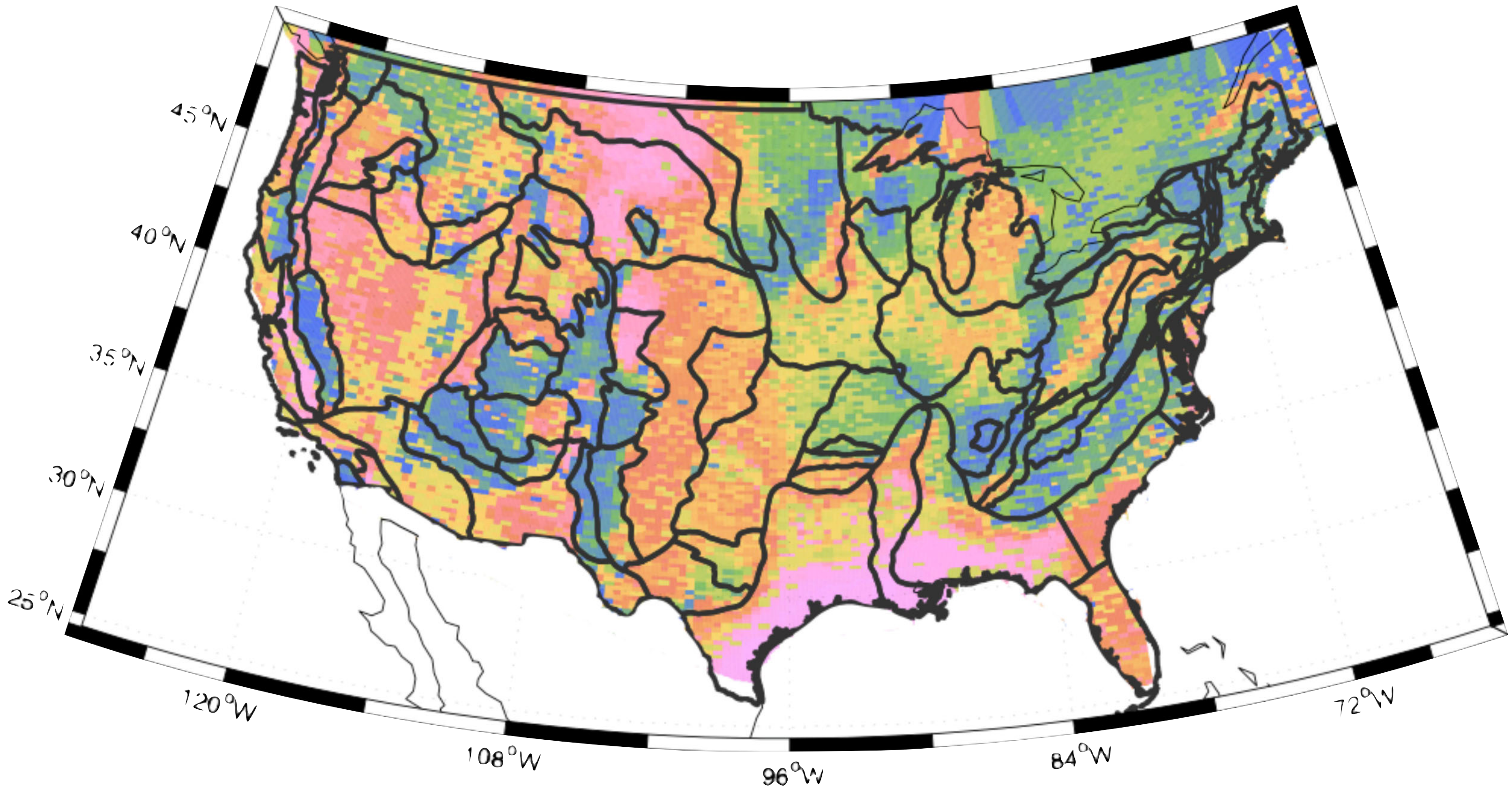
frequency

with V. Lekic (Univ. of Maryland)

Map making in seismology

surface wave velocity
ordered by theSequencer.org

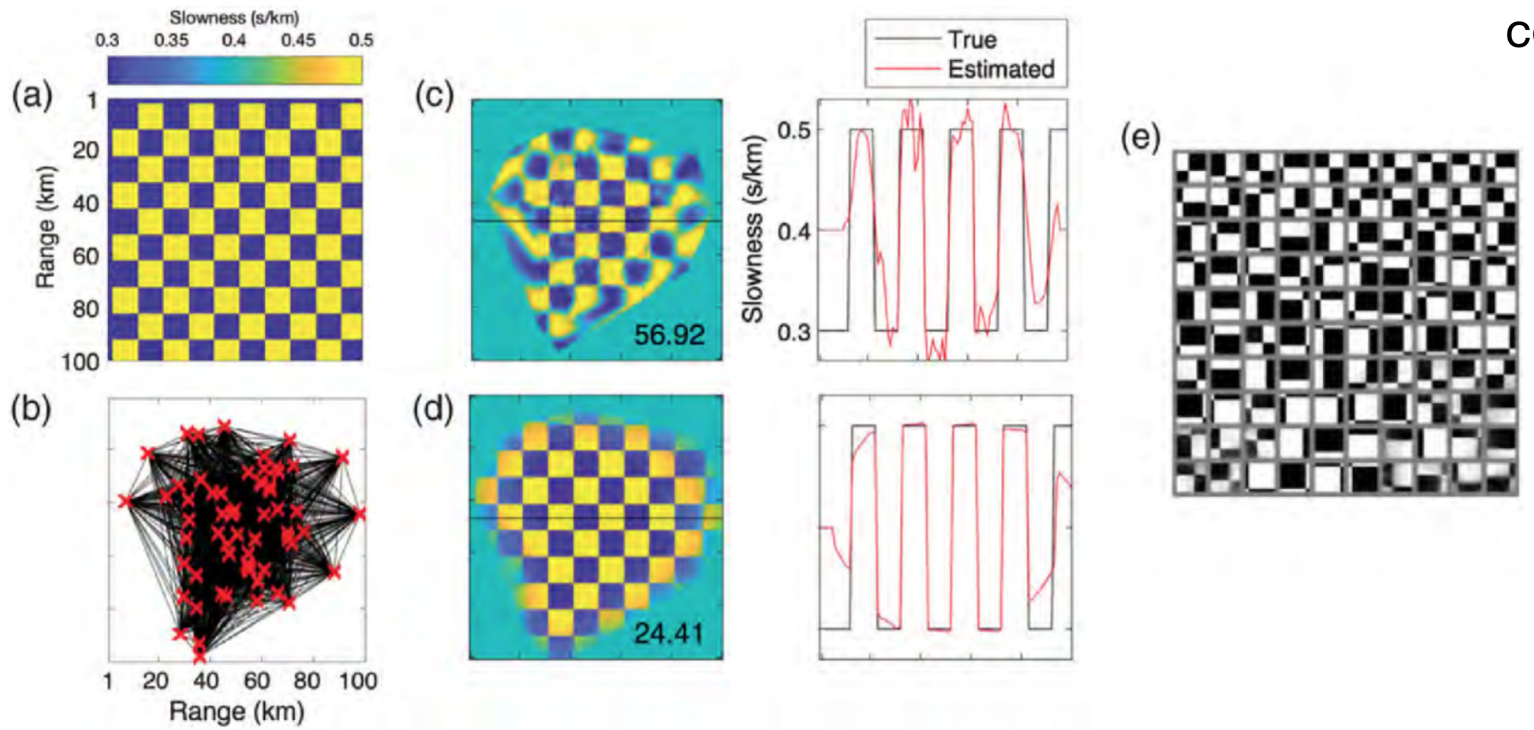




with V. Lekic (Univ. of Maryland)

Gaussian statistics, PCA — manifold learning — **dictionary learning** — (convolutional) neural nets

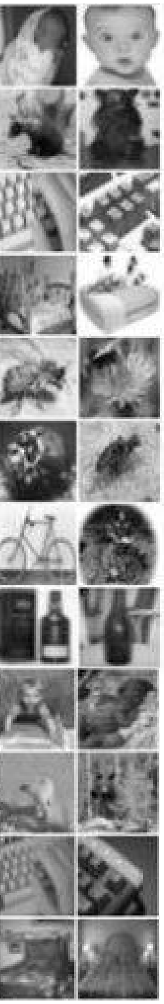
complexity

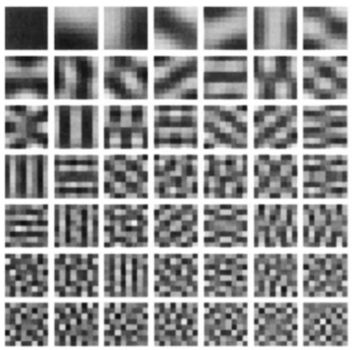


▲ **Figure 7.** Locally sparse travel-time tomography (LST; Bianco and Gerstoft, 2018) of checkerboard slowness. (a) Synthetic checkerboard slowness patterns with 100 × 100 pixel grid (km) are sampled by (b) 2016 straight rays from 64 seismic stations. (c) Conventional inversion using damping and smoothing regularization (Aster *et al.*, 2011) and (d) LST. Profiles from the 2D inversion are shown with true and estimated slownesses. The root mean square error (ms/km) estimated relative to the true slowness is printed on the 2D estimates. (e) Dictionary learned from LST contains checkerboard-like atom (100 atoms shown). Each atom (patch) is 10 × 10 pixels.

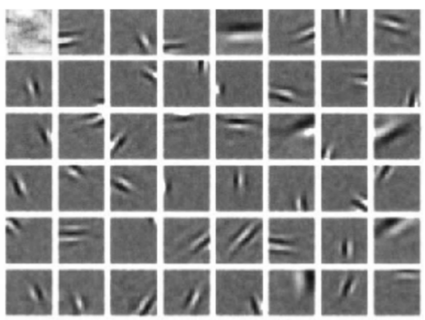
from Kong et al. (2019)

Gaussian statistics, PCA | manifold learning | **dictionary learning** | (convolutional) neural nets





Principal Component Analysis

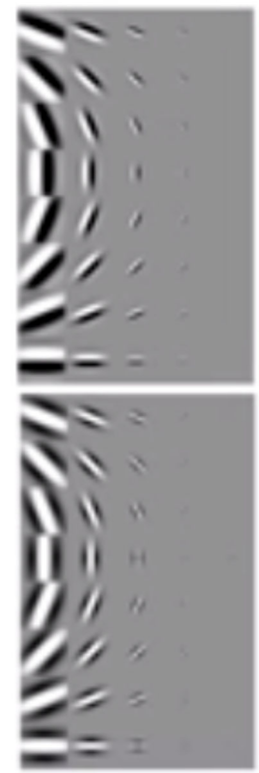


Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

kernels learned in the first layer of a deep convolutional neural net with ImageNet



wavelets (Gabor filters)

Gaussian statistics, PCA

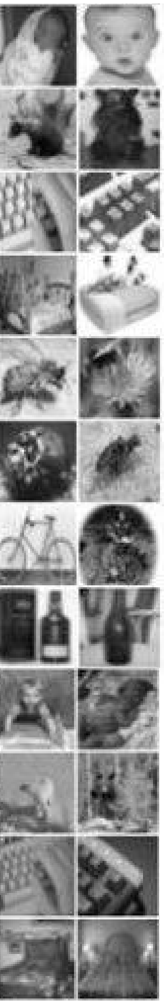
manifold learning

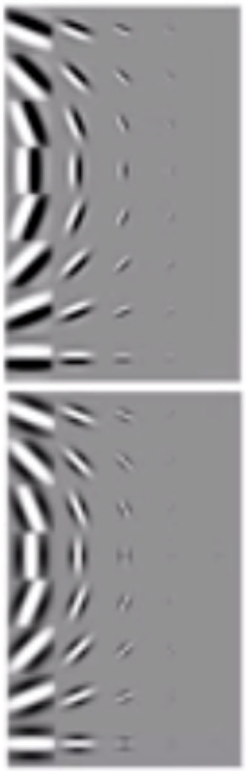
dictionary learning

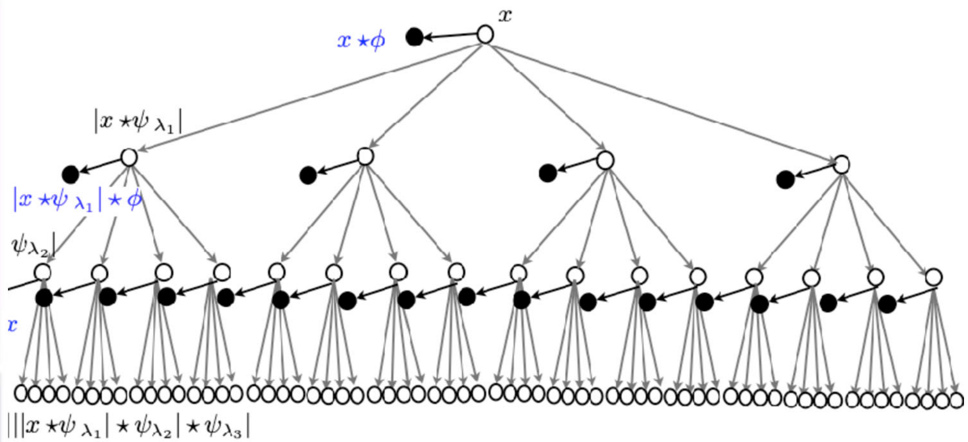
(convolutional) neural nets

complexity

The scattering transform



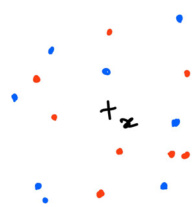


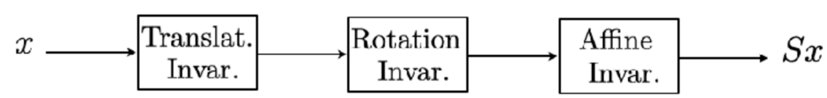


Invariant to:

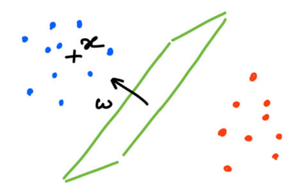
- translations
- rotations
- small deformations

Data $x \in \mathbb{R}^d$



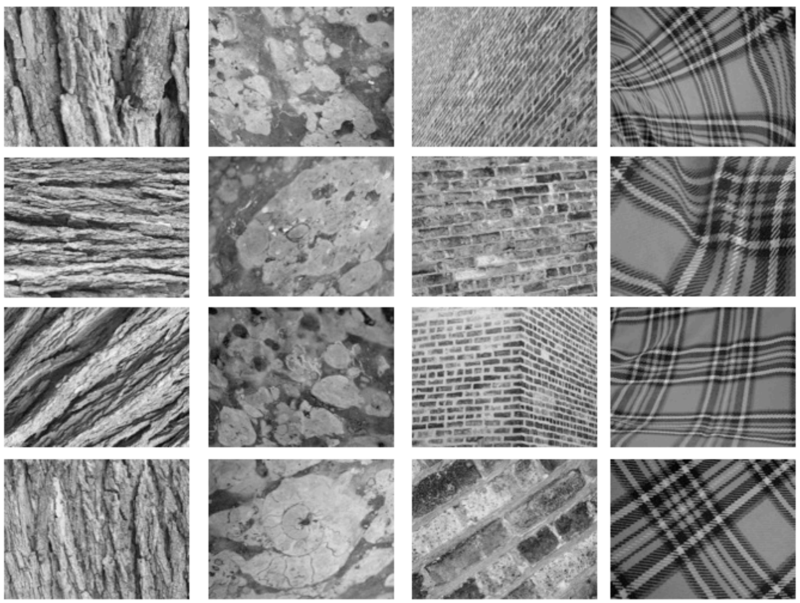


Representation

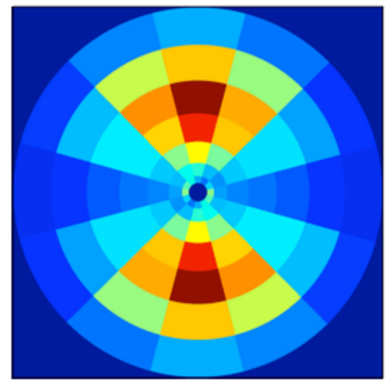


J. Bruna & S. Mallat

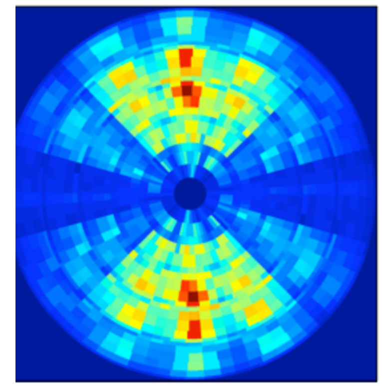
Gaussian statistics, PCA → manifold learning → **dictionary learning** → (convolutional) neural nets



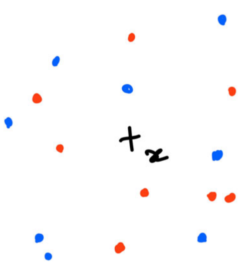
first order coeffs



second order

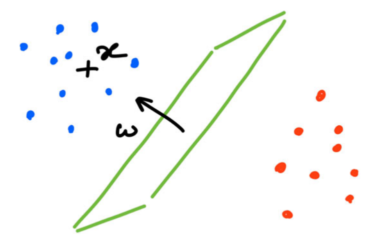


Data $x \in \mathbb{R}^d$



$$x \longrightarrow \boxed{\begin{array}{c}\text{Translat.}\\ \text{Invar.}\end{array}} \longrightarrow \boxed{\begin{array}{c}\text{Rotation}\\ \text{Invar.}\end{array}} \longrightarrow \boxed{\begin{array}{c}\text{Affine}\\ \text{Invar.}\end{array}} \longrightarrow Sx$$

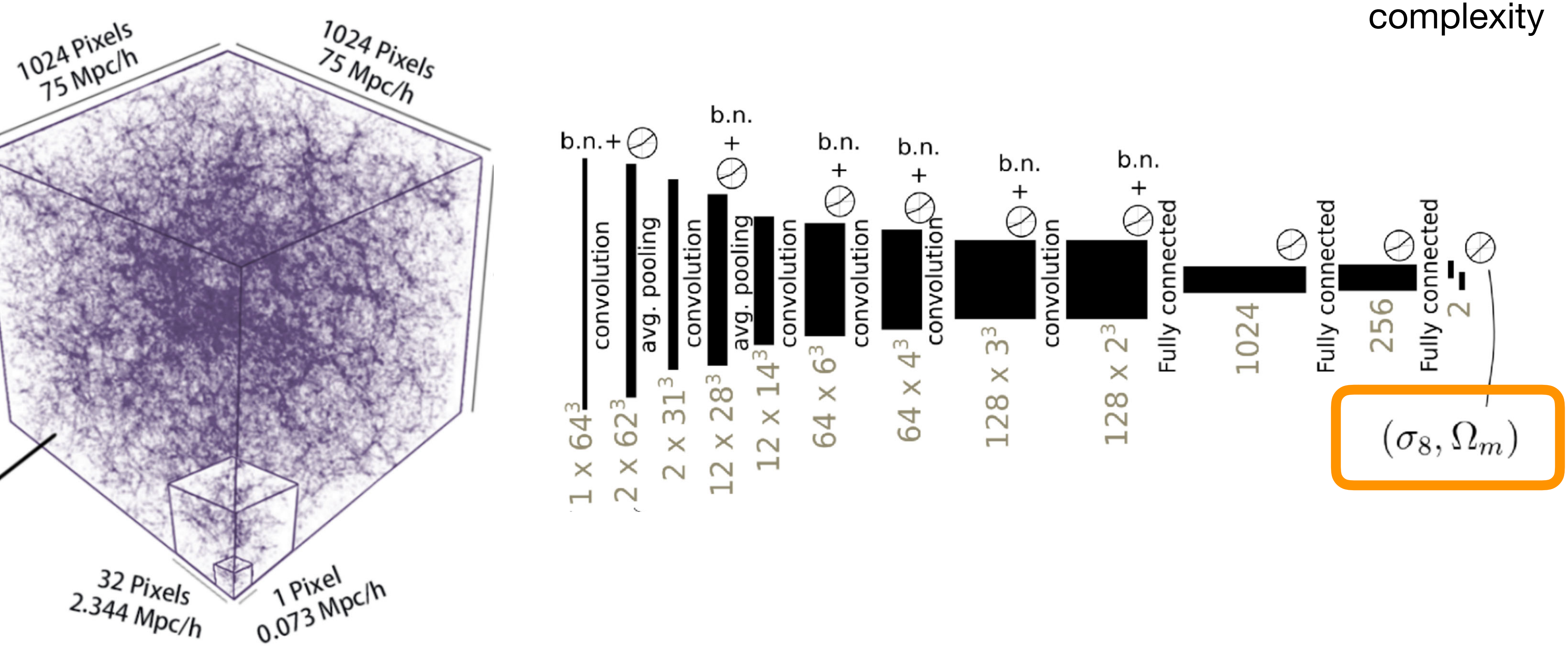
Representation



from Laurent Sifre & S. Mallat

Gaussian statistics | manifold learning | dictionary learning | **(convolutional) neural nets**

complexity



Ravanbakhsh et al. (2016)

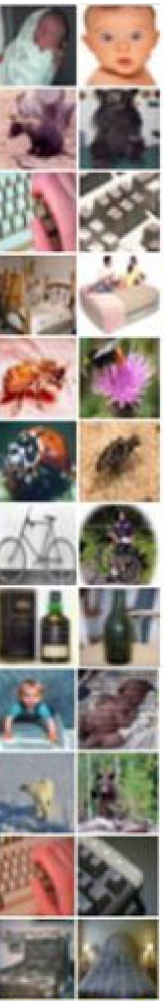
Gaussian statistics

manifold learning

dictionary learning

(convolutional) neural nets

complexity



dictionary

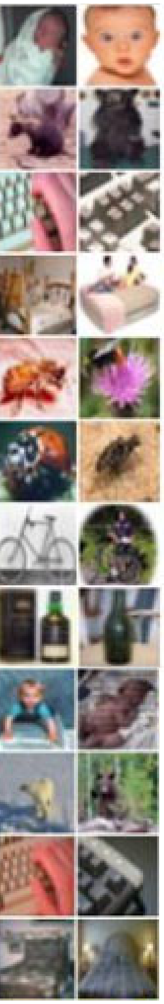
Gaussian statistics, PCA

manifold learning

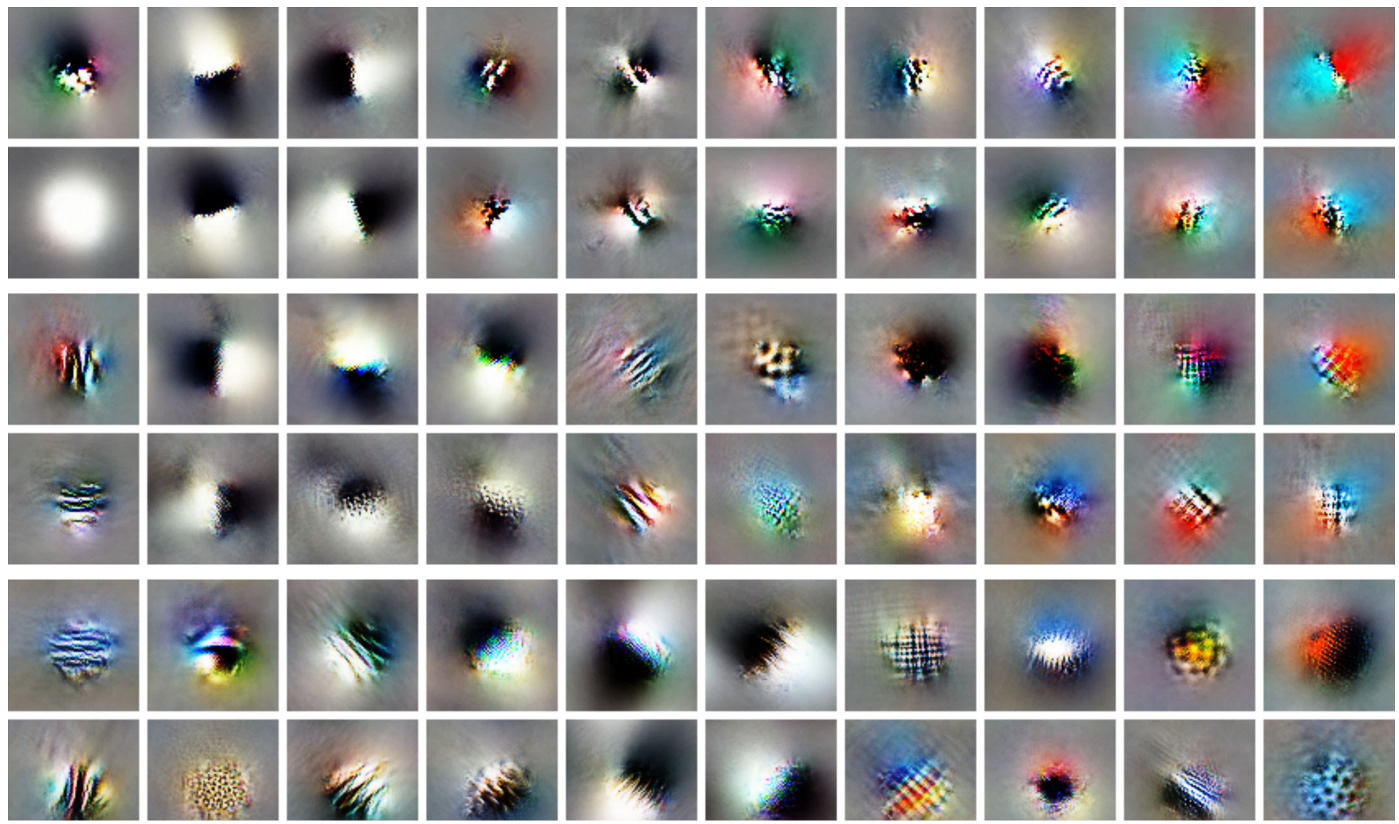
dictionary learning

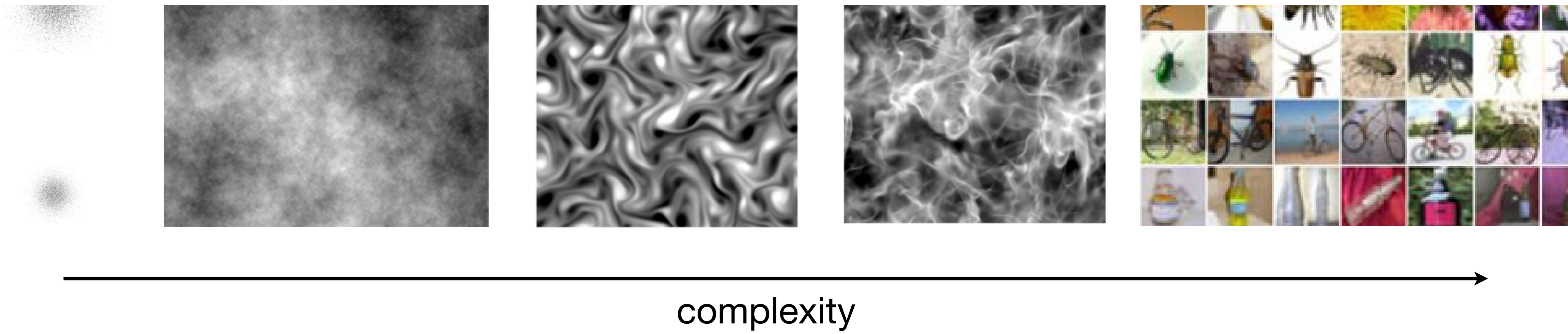
(convolutional) neural nets

complexity



collaboration with OpenAI

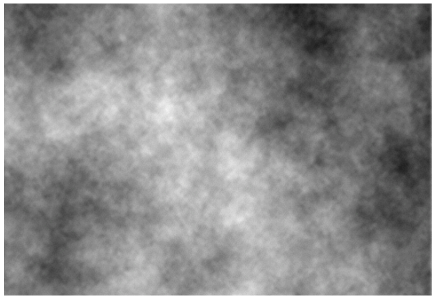
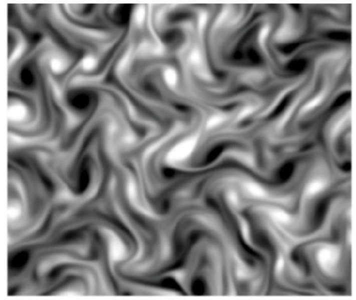
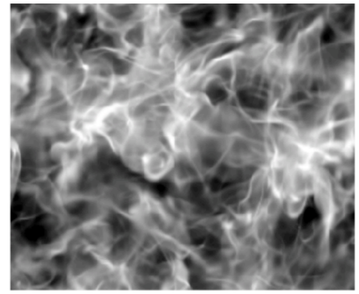




As scientists, we want to describe/explain the complex world

We want to go from complexity to simplicity.

This requires an efficient language.

complexity

Gaussian statistics	manifold learning	dictionary learning	(convolutional) neural nets

Huge space to explore

Simplicity found in:

parameters	geometry	vocabulary	classes

Challenges: interpretability, control of systematics