Renormalized Mutual Information for Artificial Scientific Discovery

Leopoldo Sarra

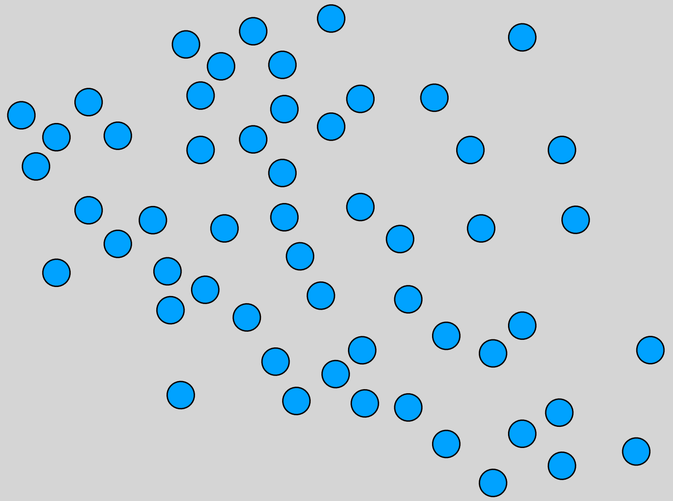
Max Planck Institute for the Science of Light
Erlangen, Germany

work with **Andrea Aiello** and **Florian Marquardt**

Learning to Discover - AI and Physics Conference
Thursday, April 26th 2022

MAX PLANCK INSTITUTE
for the science of light

How can we describe a physical system with only a few quantities?



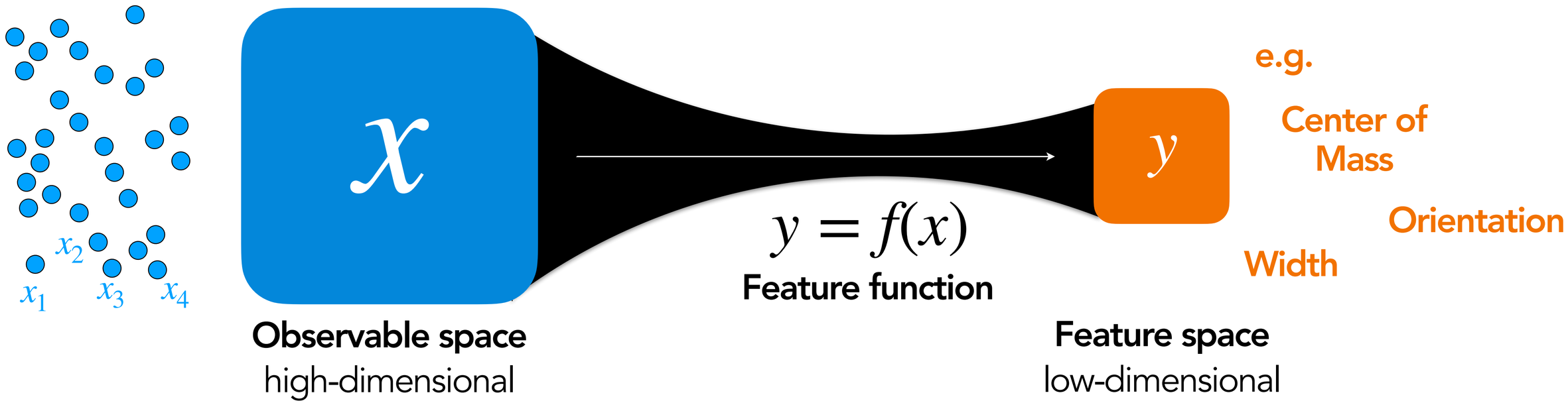
Center of Mass

Orientation of the cloud

Width of the cloud

e. g. Fluid dynamics, Thermodynamics, …

Features

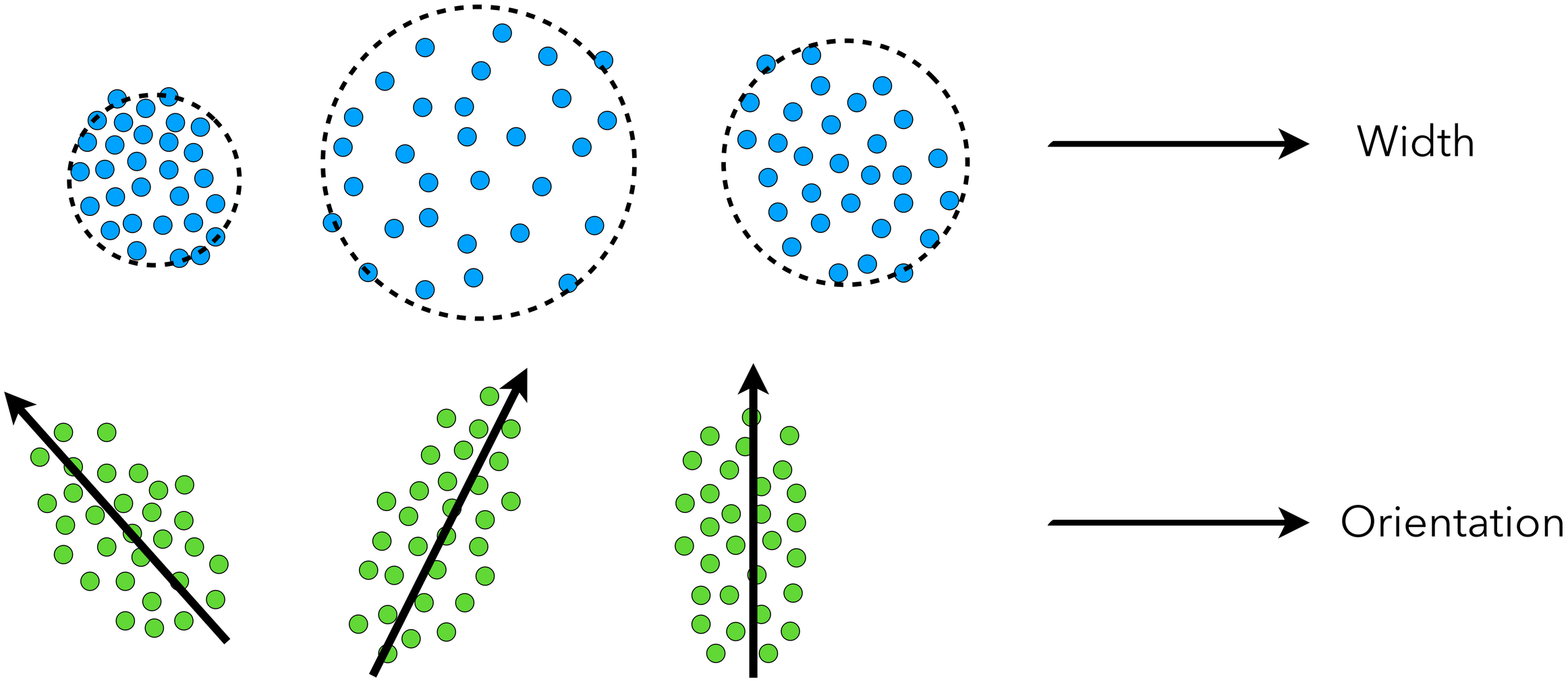


Goal: Given a high-dimensional system x, we want to automatically find the best low-dimensional features $y = f(x)$ to describe it.

References: Haykin, Simon S. Neural Networks: A Comprehensive Foundation. 2nd ed. Upper Saddle River, N.J: Prentice Hall, 1999.

Feature Extraction

Statistical analysis of many observations



Mutual Information

If two random variables x and y are dependent one another, and we are provided with the value of y, how much do we learn on x?

Entropy

$$I(x,y) = H(y) - H(y\,|\,x) = \int dxdyP(x,y)\log\frac{P(x,y)}{P_x(x)P_y(y)}$$

Conditional Entropy

- It quantifies the dependence between two random variables
- Always positive, zero iff the variables are independent

Feature Extraction: choose $f(x)$ such that $I(x, y = f(x))$ is maximized

References:
- *Mutual Information - e. g. Papoulis, Athanasios, and S. Unnikrishna Pillai. Probability, Random Variables, and Stochastic Processes*
- *Bell, Sejnowsky. "InfoMax: An Information-maximisation Approach to Blind Separation and Blind Deconvolution";*

Problem:

Mutual Information is $+\infty$ for any feature!

$$H(y\,|\,x) = -\int dxdy P_x(x) P(y\,|\,x) \log P(y\,|\,x)$$

A continuous deterministic feature $y = f(x)$ has $P(y\,|\,x) = \delta(y - f(x))$.

$$H(y = f(x)\,|\,x) = -\int dxdy P_x(x) \delta(y - f(x)) \log \delta(y - f(x)) = -\log \delta(0)$$

it always diverges in this case!

$$-\infty$$

Long-standing problem but unsatisfying solutions

References:

- *Bell, Sejnowsky. "InfoMax: An Information-maximisation Approach to Blind Separation and Blind Deconvolution";*
- *Gabrié et al."Entropy and Mutual Information in Models of Deep Neural Networks." Journal of Statistical Mechanics;*
- *Saxe et al."On the Information Bottleneck Theory of Deep Learning"*

Renormalized Mutual Information

our solution

We add Gaussian noise λ to the x and define a new finite quantity

$$I_\varepsilon(x, y) = I\left(x, y = f(x + \varepsilon\lambda)\right)$$

We perform the zero-noise limit:

$$\tilde{I}(x, y) = \lim_{\varepsilon \to 0} I_\varepsilon(x, y) + H(\varepsilon\lambda) = H(y) - \int dx P_x(x) \log \sqrt{|\nabla f(x) \cdot \nabla f(x)|}$$

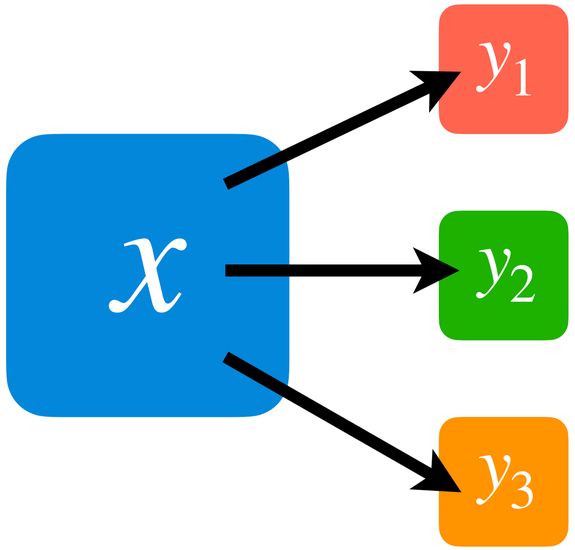
- Finite quantity and well-defined
- Invariant under feature reparametrization
  (an invertible transformation $y' = g(y)$ does not change $\tilde{I}$)

L. Sarra, A. Aiello, F. Marquardt

arXiv:2005.01912

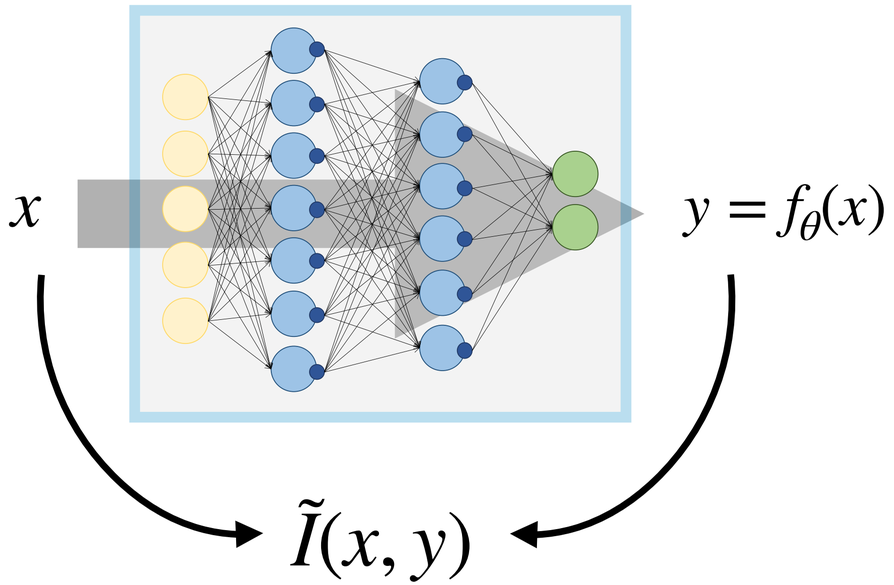
Applications

Feature Selection



Find out how useful a given macroscopic quantity is to describe the system

Feature Extraction

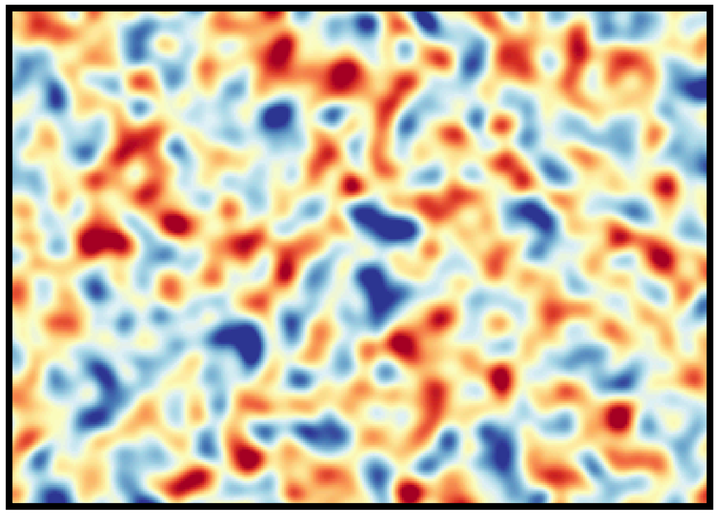


Optimize over a class of functions parametrized by a neural network

Possible Applications

- Characterize different phases of matter
- Partial prediction of time evolution
- Discover residual regularities of chaotic systems
- Study of intermediate layers of neural networks during training
- Unsupervised representation learning
- …

Fluctuating fields

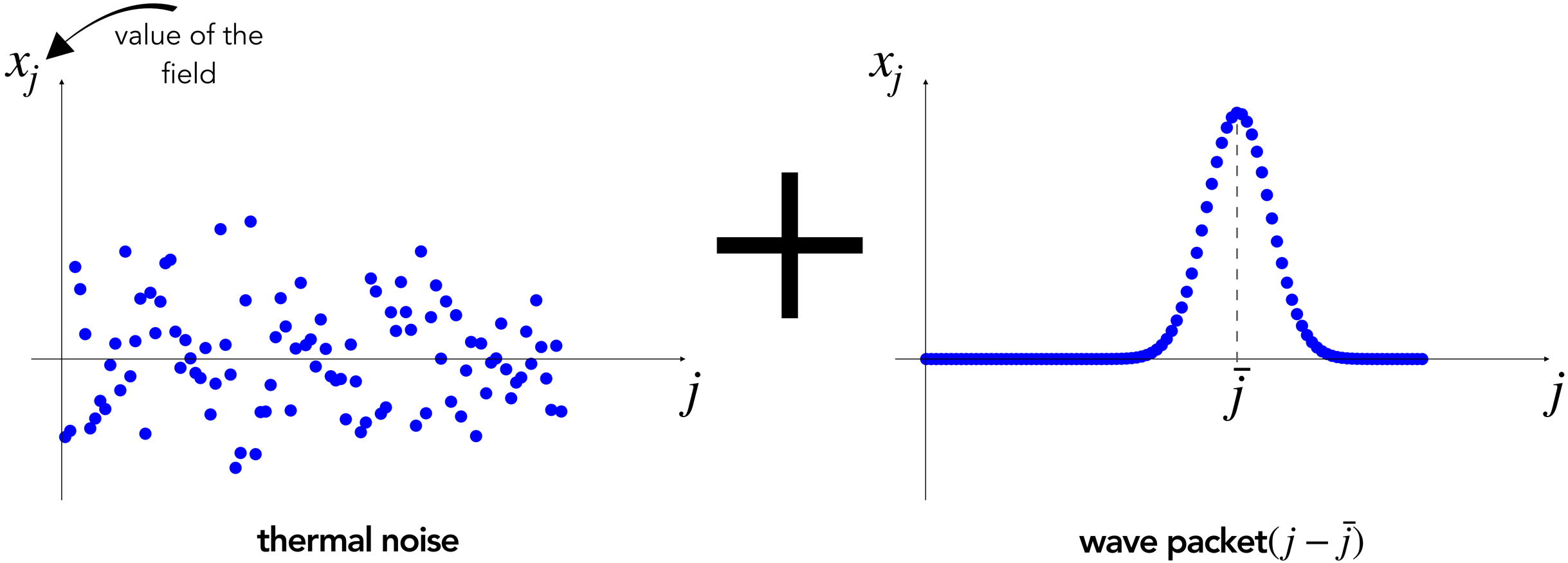


Many-particle systems

ESA/Hubble

EXAMPLE

Fluctuating field with wave packet

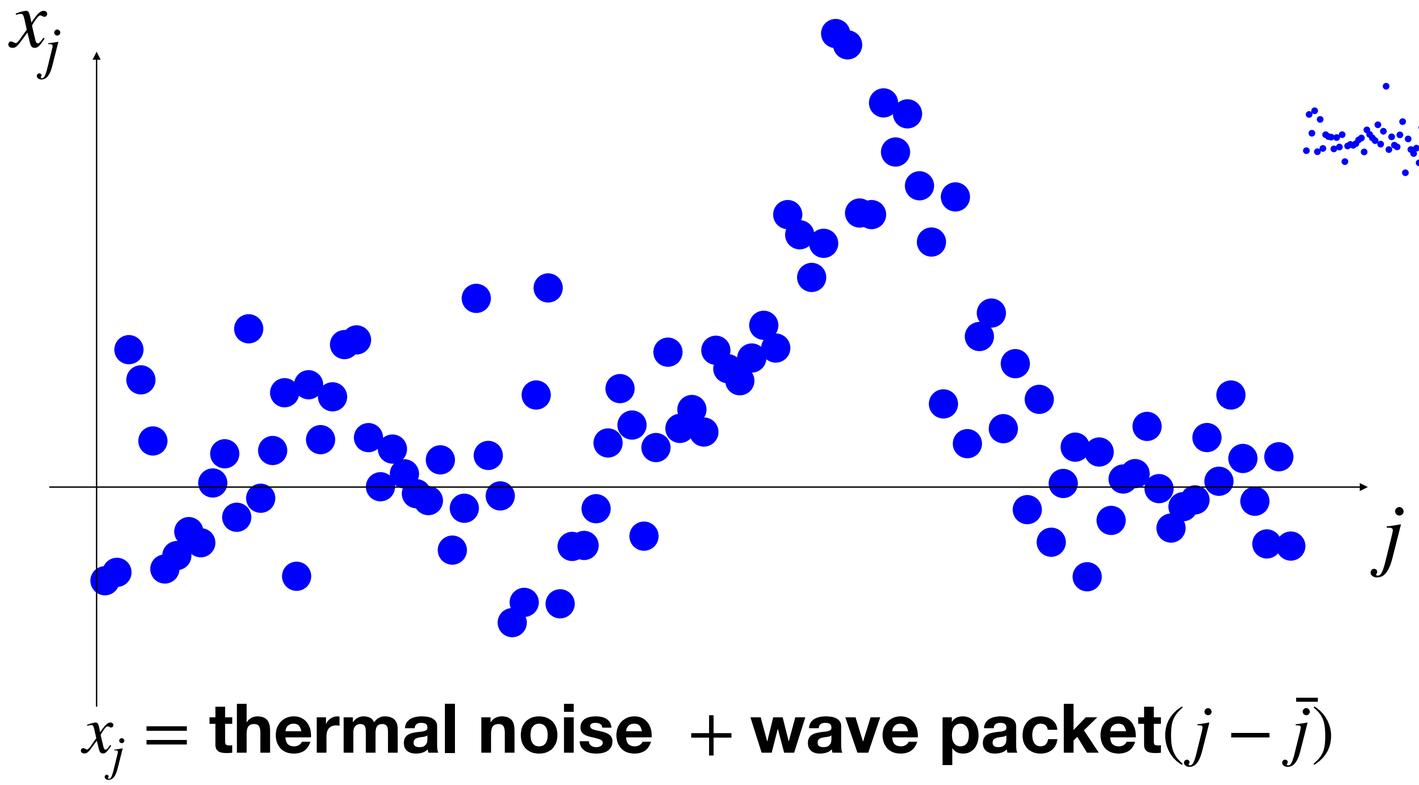


thermal noise

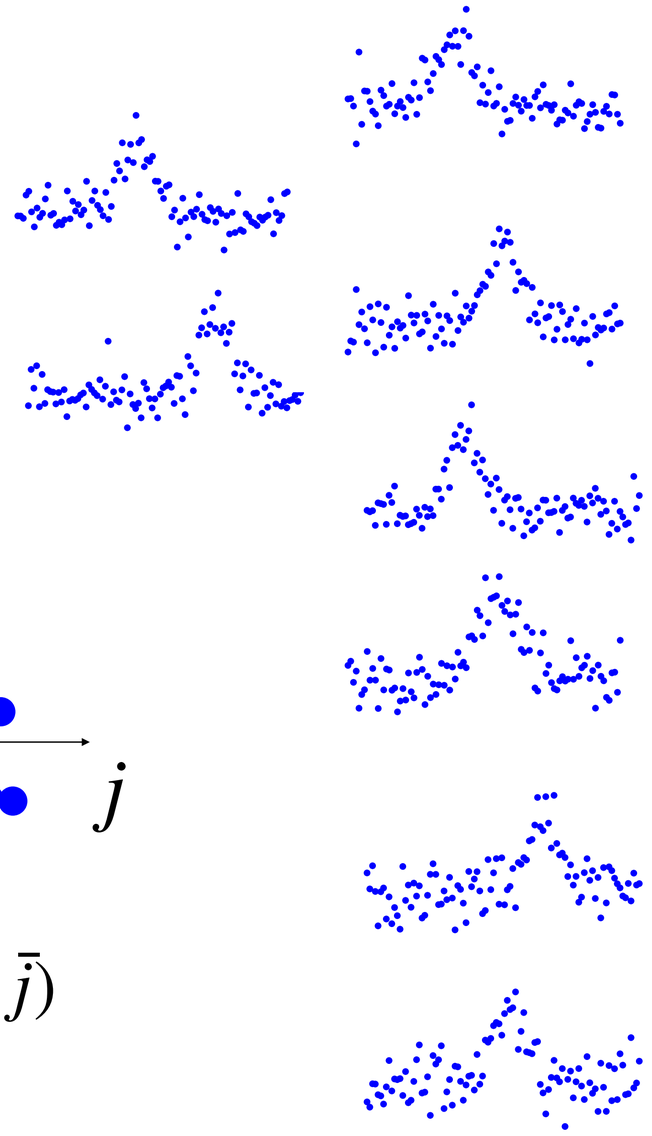
wave packet$(j - \bar{j})$

EXAMPLE

Fluctuating field with wave packet



x_j = **thermal noise** + **wave packet**$(j - \bar{j})$



EXAMPLE

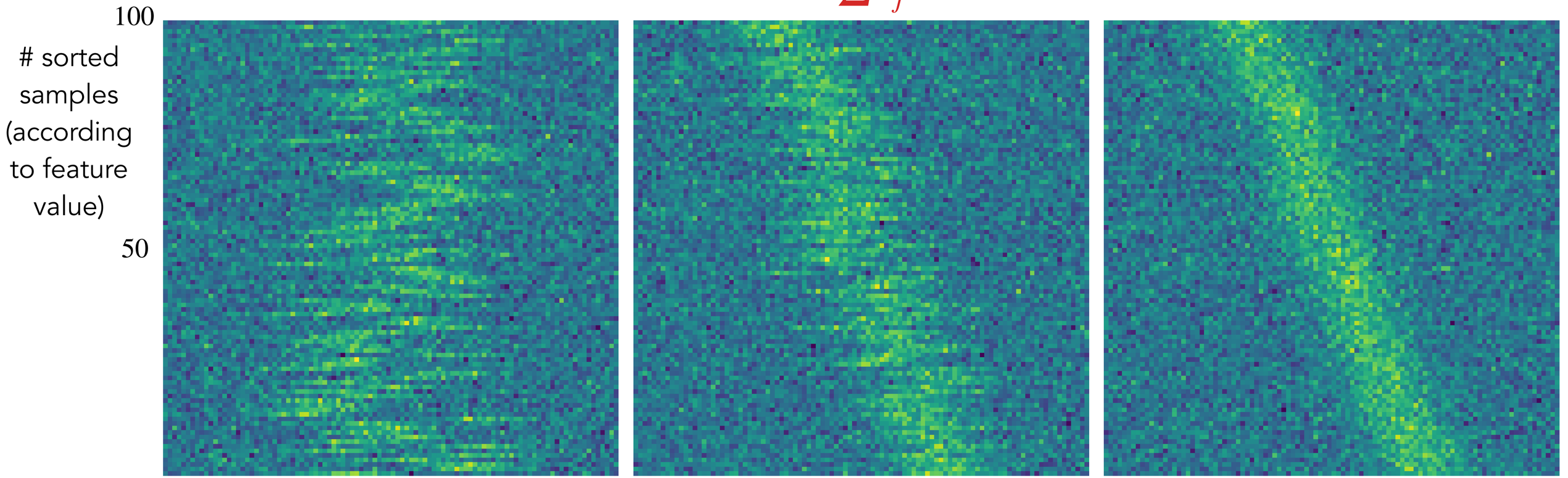
Fluctuating field with wave packet

x_j

$$y = \sum x_j^2$$

$$y = \frac{\sum j x_j^2}{\sum x_j^2}$$

$$y = f_{NN}(x)$$



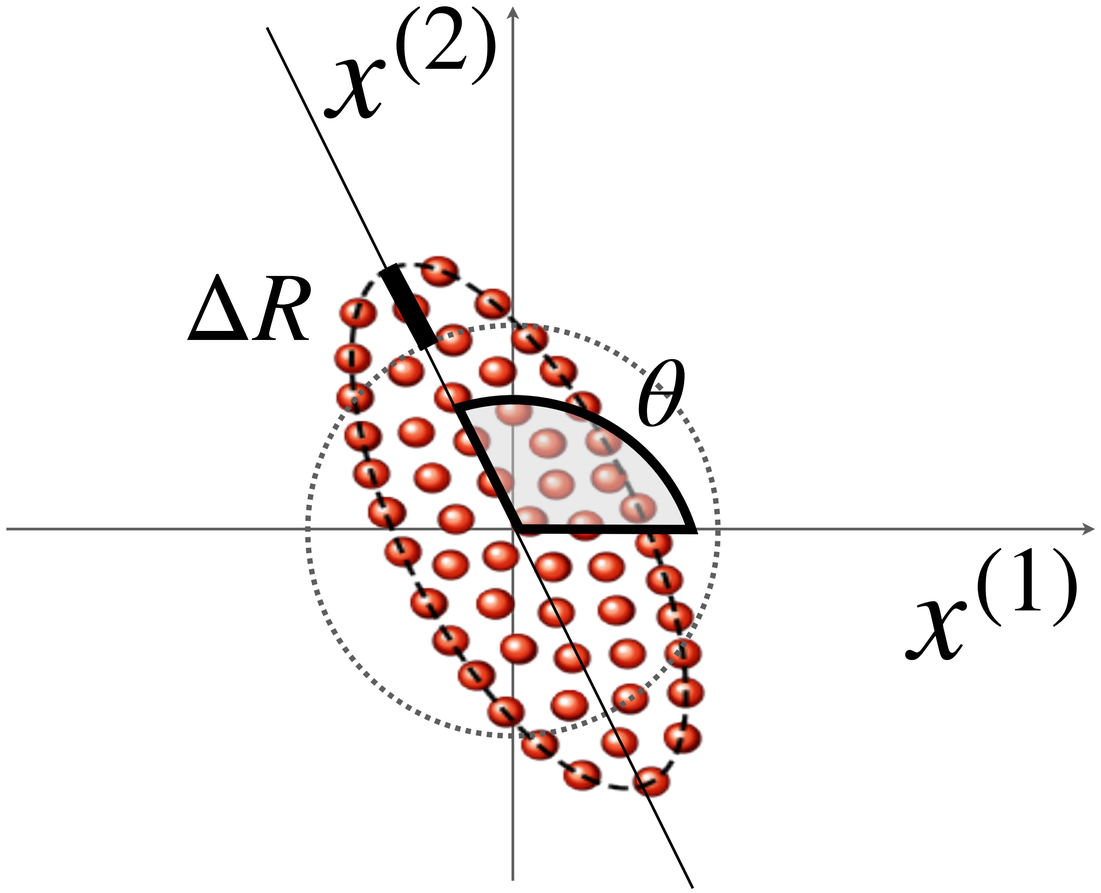
Renormalized Mutual Information:

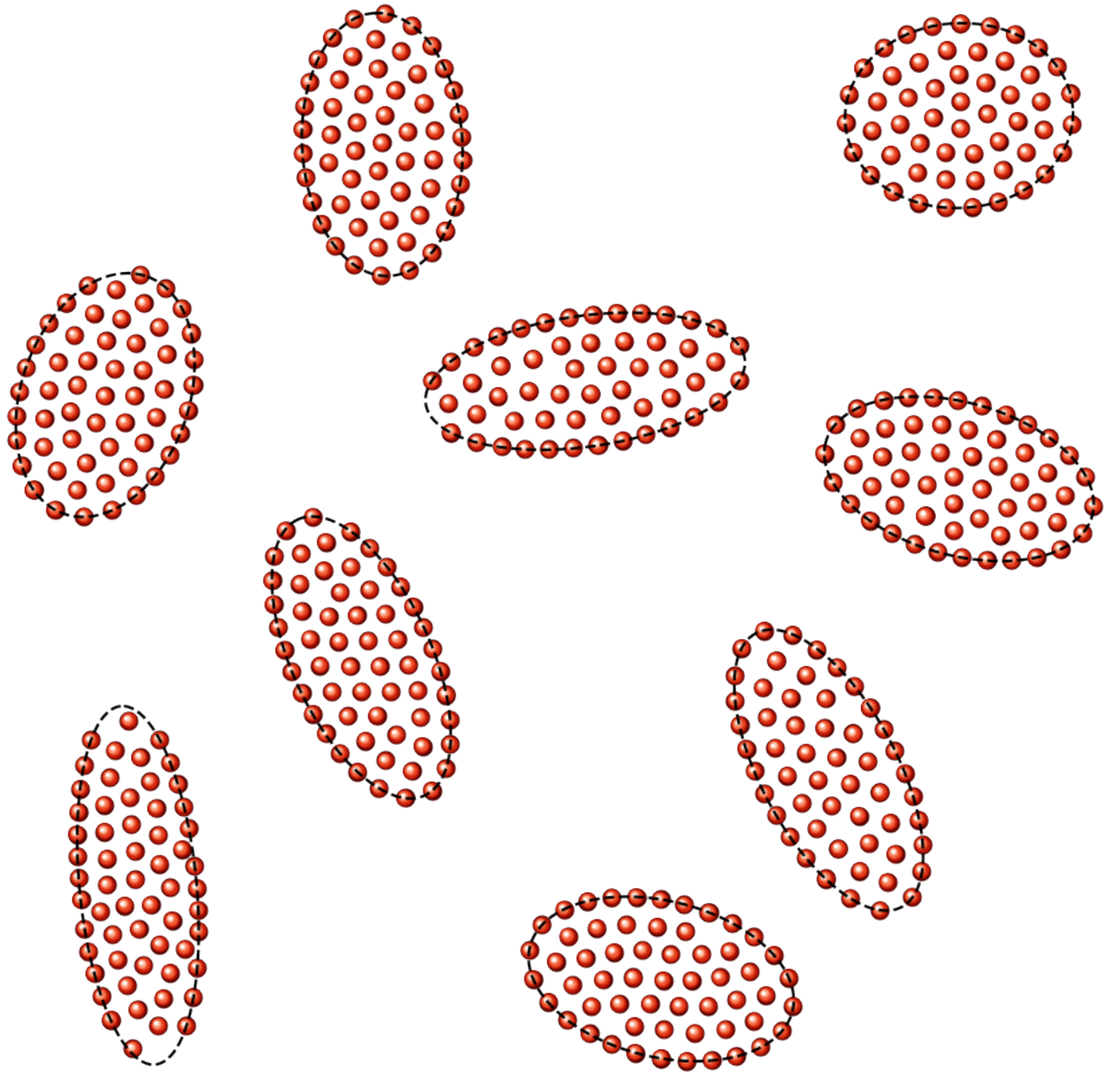
0.29 | 0.93 | 2.73

Reference: in our paper *arXiv:2005.01912*, we also estimate the quality of the feature representations in a more quantitative way by comparing the performance on a supervised regression task

EXAMPLE

Liquid Drop





Each ellipse has a different
deformation and **orientation**

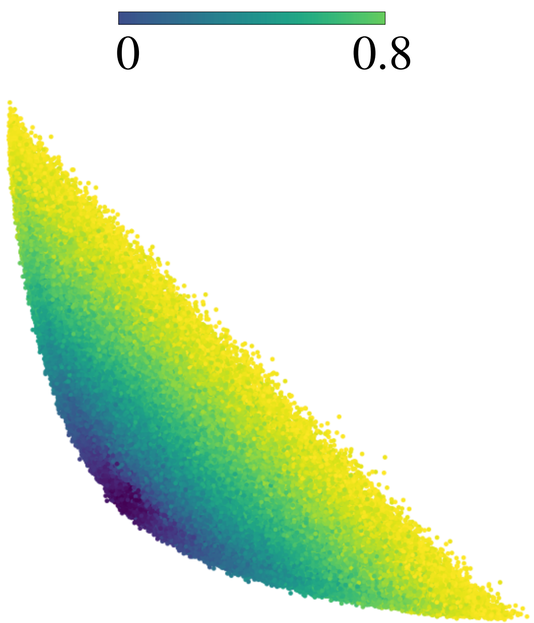
EXAMPLE

Liquid Drop

Handcrafted Features

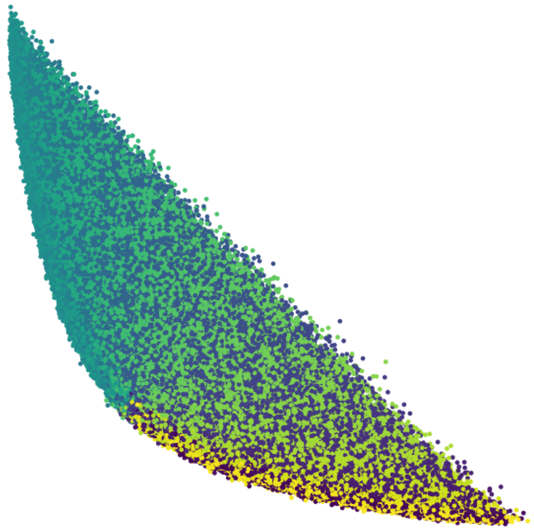
Δr

0 0.8



θ

0 π



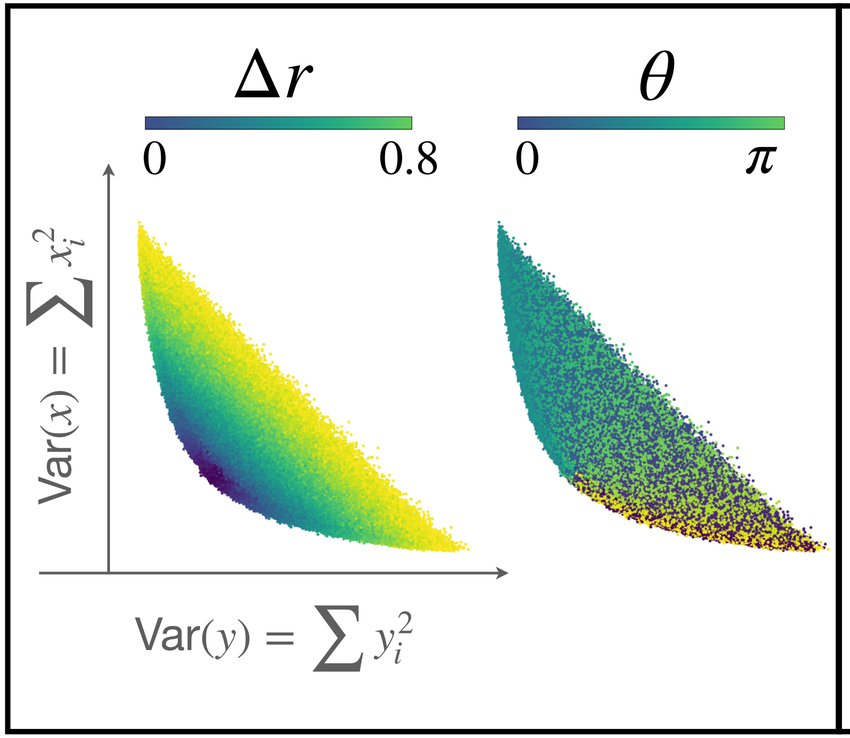
$\text{Var}(x) = \sum y_i^2$

$\text{Var}(y) = \sum x_i^2$

EXAMPLE

Liquid Drop

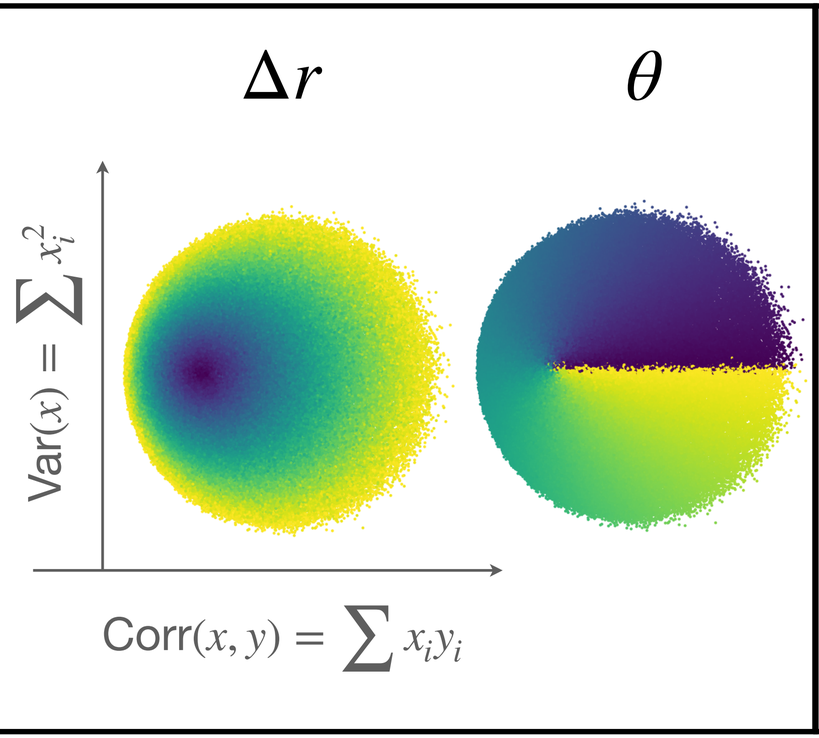
Handcrafted Feature



Renormalized Mutual Information:

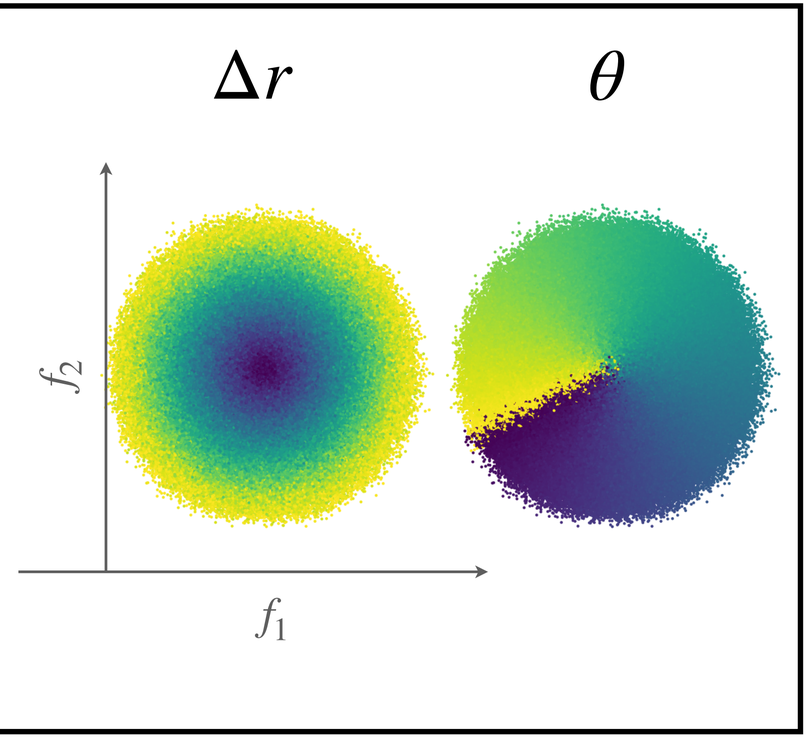
1.75

Handcrafted Feature



3.04

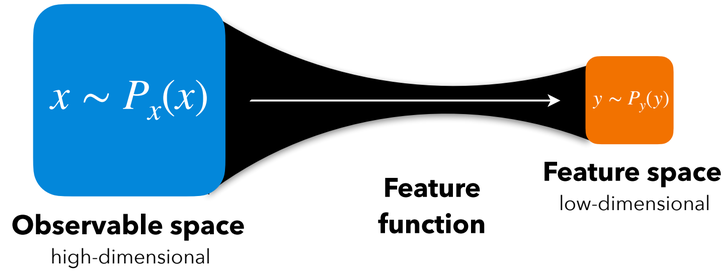
Extracted Feature (neural network)



3.22

Reference: in our paper *arXiv:2005.01912*, we also estimate the quality of the feature representations in a more quantitative way by comparing the performance on a supervised regression task

Outlook



What are the most relevant features of a system?

$$\tilde{I}(x,y) = H(y) - \int dx P_x(x) \log \sqrt{|\nabla f(x) \cdot \nabla f(x)|}$$

Many possible applications in Physics and Machine Learning

- Characterize different phases of matter
- Partial prediction of time evolution
- Discover residual regularities of chaotic systems
- Study of intermediate layers of neural networks during training
- Unsupervised representation learning
- …

"*Renormalized Mutual Information for Artificial Scientific Discovery*",
LS, Andrea Aiello, and Florian Marquardt,
Phys. Rev. Lett. **126**, 200601