Jakovác Antal

WIGNER

Dept. of Computational Sciences

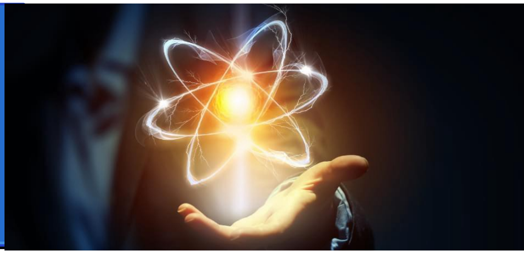
Entropy of (artificial) intelligence

PP 2022, Budapest, Margaret Island, May 15-18, 2022.

Outlines

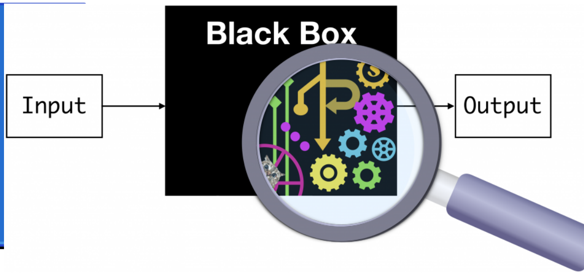
- Introduction, motivation
- Examples of data modeling
- Mathematical formulation
- Number of relevant features
- Entropy of intelligence
- Conclusions

Motivation: a new scientific paradigm



- **scientific approach**: observation, mathematical description, solution and predictions
- **traditionally**: everything is done by hand
  - set of solvable models is very limited (mostly linear systems) ⇨ influence worldview
- **from mid XX. century**: solution by computers
  - much more models to solve ⇨ new disciplines (chaotic systems, MC simulations)
  - meaning of the results? *new fixed points, we can compute, but we do not understand*
  - significance of the terms in the equations? *cf. meaning of terms in the Lagrangian*
- **XXI. century**: numerical determination of mathematical description (equations)
  - much more phenomena can be studied – aesthetics? interpretability?
  - task of humans: symmetries, educated guess of laws

Understanding understanding



We have to understand, how we build a model for a problem.

Some guiding principles:

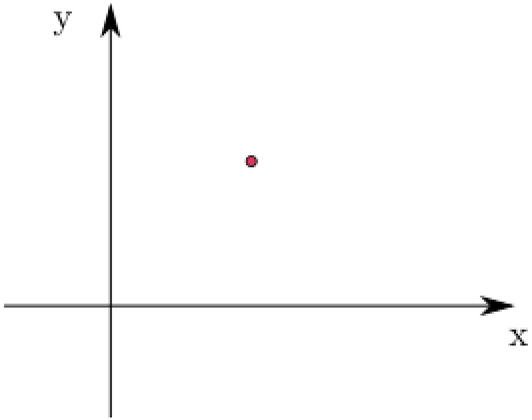
- all information must be present in the data (*data-driven modeling*)
- *unsupervised* learning: no human action is required beyond to present the system to be understood
- humans make models by observing different adequate *features* (eg. a solid object, animal, has two legs, wings, wolf-like muzzle and ears, etc.).
  Features → measurement, characterization, coordination
- **understanding ≡ best representation of data**
  (finding the proper coordinate system to describe data)

(remark: this narrative is different from the one behind supervised learning; we have no information loss, only different representation)

Examples of data modeling

The most elementary, but generic task is to tell if an item is element of a set.

Continuous examples: single 2D data point: S={p} one element set.



We can represent it with the (x,y) coordinates.

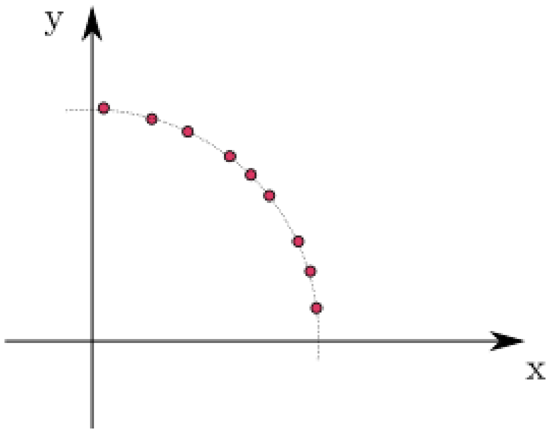
Other representations are also appropriate.

For a single data all representations are equivalent.

Examples of data modeling

The most elementary, but generic task is to tell if an item is element of a set.

Continuous examples: multiple 2D data points



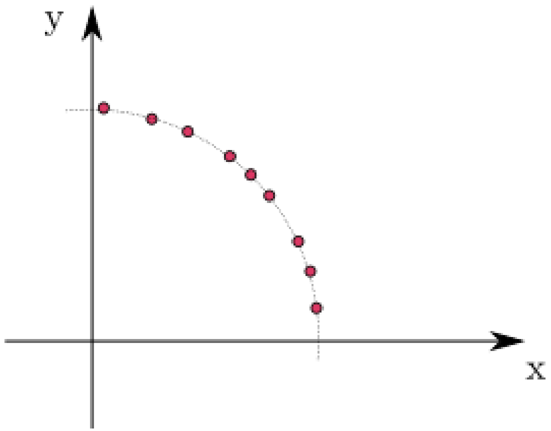
In the (x,y) representation the coordinates are not independent.

In the polar coordinate system (r,φ) we find r=R for all data points! The r and φ coordinates are independent.

Examples of data modeling

The most elementary, but generic task is to tell if an item is element of a set.

Continuous examples: multiple 2D data points



In the (x,y) representation the coordinates are not independent.

In the polar coordinate system (r,φ) we find r=R for all data points! The r and φ coordinates are independent.

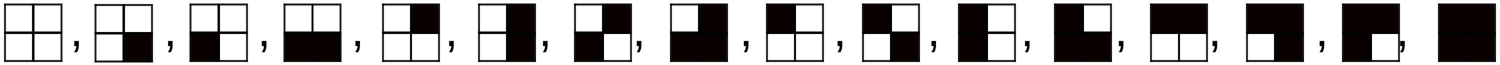
In a well-chosen coordinate system the data coordinates are independent, and they are either constant (**relevant** or **selective** coordinates, or laws),
or variable (**irrelevant** or **descriptive** coordinates).

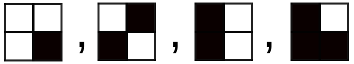
Examples of data modeling

The most elementary, but generic task is to tell if an item is element of a set.

Discrete examples: consider *2x2 bitmap “images”*, and choose a subset. Can we find the proper representation of the set where the identification of the subset is easy?

We can list all images:

X = {  }

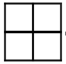
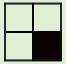
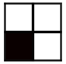
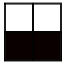
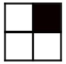
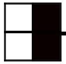
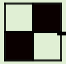
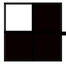
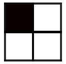
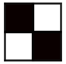
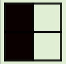
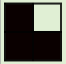
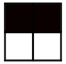
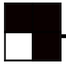
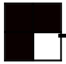
choose an arbitrary subset, our abstract “cat images”: C={  }

- the pixel-wise coordination C={0001,0110,1010,1011} : no regularity
- the pixels are not independent in C:

$$P(\xi_1=0,\xi_2=0)=1/4 \neq P(\xi_1=0)P(\xi_2=0)=1/2*3/4$$

Examples of data modeling

Find a coordination that fits the best to the problem!

X = {→0100, →0000, →0101, →0110, →0111, →1000, →0001, →1001,
→1010, →1011, →0010, →0011, →1100, →1101, →1110, →1111}

This is *not the original bit coordinates,* but it fits well to our chosen C subset!

In the new coordinates: C = {0000,0001,0010,0011}

- first two bits are 0 for elements of C: these are the relevant (selective) coordinates:
  $x \in C \Leftrightarrow x_0 = x_1 = 0$:appropriate to select the elements of C
- last two bits are variable: these are the irrelevant (descriptive) coordinates:
  to tell apart elements of C (compression) we need to consider only these coordinates

Coordination and understanding

Features: independent coordinates over C, either selective or descriptive

Let ξ be the common features for $C_{1,}C_{2,}\ldots,C_a, C=\cup_i C_i$

- **classification:** $x\in C_i$ iff selective bits of $\xi(x)=$selective bits of C_i
- **decoding:** to produce $x\in C_i$ we have to chose the relevant bits characteristic to C_i and the irrelevant bits independently, uniform randomly

$$\xi^{-1}(\sigma_{relevant}=C_{i,relevant}, \sigma_{irrelevant}=\text{random})\in C_i$$

- **lossless data compression:** if we know that $x\in C_i$, the relevant bits can be built into the static part of the code, and we have to store the *irrelevant bits*.

All the AI tasks can be solved by inspecting certain bits.

Mathematical description

- **basic approach**: dual interpretation of a coordination (physical quantities)
  - we want to describe a set X (e.g. a glass of water, or 1Mpixel images, assumed to be finite)
  - coordination: $\xi = (\xi_1, \ldots, \xi_N)$, $\xi : X \rightarrow B^N$ $(B \subset \mathbb{R})$ bijection
  - random variables: $X \rightarrow B$ $(B \subset \mathbb{R})$ functions
- coordinates can be interpreted as random variables, too! We can define the distribution, or the statistical independence of the coordinates.
- **measure space** is (X, F_X, P_X) where $F_X = 2^X$ and $P_X(C \subset X) = \frac{|C|}{|X|}$
- *joint distribution* of $I = i_1, i_2, \ldots, i_a$ components of ξ over $C \subset X$ is a conditional probability

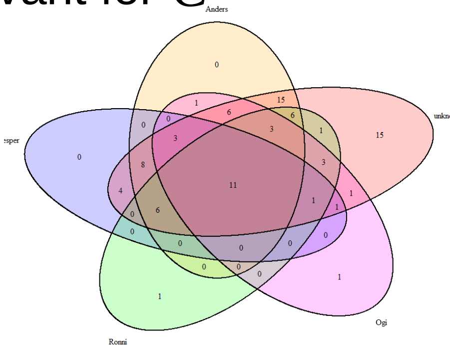
$$p_C(\xi_I = \sigma_I) = P_X(\{x \in C \mid \xi_i = \sigma_i \;\forall i \in I\}) = \frac{|\xi_I^{-1}(\sigma_I) \cap C|}{|C|}$$

Existence of complete feature set

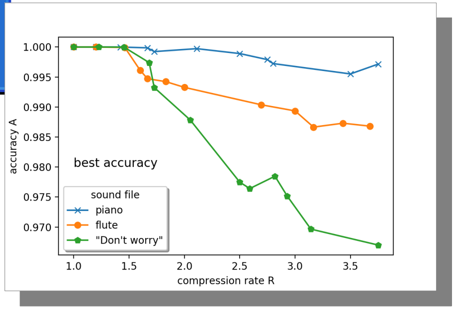
Statement:

In every $C_{1,}\ldots,C_a,C\subset X$ subsets, where C_i are pairwise disjoint and $\cup C_i=C$, we can define a $\xi:\bar{X}\rightarrow B^N$ bijection over extended $\bar{X}\subset X\times B$, where the components are independent over extended $\bar{C}\subset C\times B$, and they are either relevant or irrelevant with respect to all extended $\bar{C}_{1,}\ldots,\bar{C}_a,\bar{C}$ $(\bar{C}_i\subset C_i\times B)$.

- *overall relevant/selective features:* relevant coordinates of C
- *partially relevant/selective features:* relevant for some C_i , but irrelevant for C
- *irrelevant/descriptive features:* irrelevant for all C_i

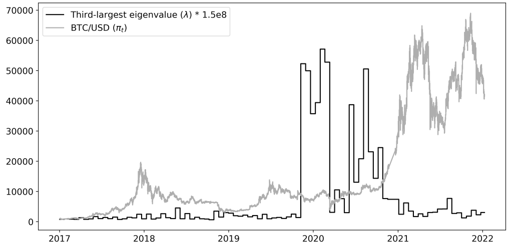


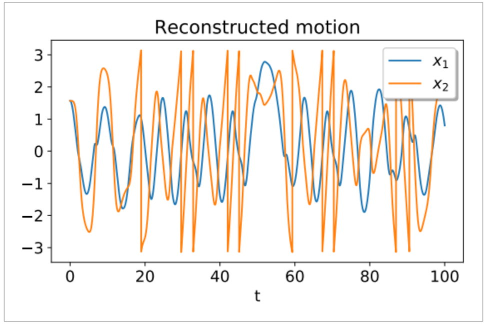
Publications in the topic



Using this technique we studied some topics:

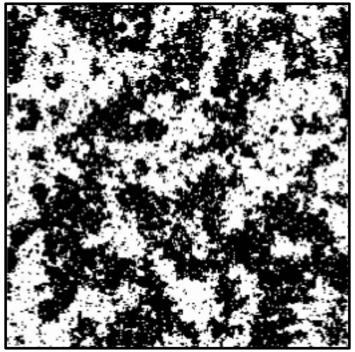
- [D.Berenyi, AJ, P. Pósfay, 2020]: paper about the theoretical basics
- [AJ, 2021]: treating linear laws, application for musical data compression
- [TS. Biró, AJ, 2022] : entropy associated to representations
- [M. Kurbucz, P. Pósfay, AJ, 2022] using linear laws we examined Bitcoin prices and identified potential external influence
- [M. Kurbucz, P. Pósfay, AJ, 2022]: reconstruction of mechanical motions using nonlinear laws
- … more in preparation



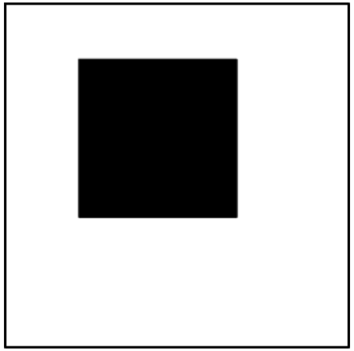


Number of relevant features

Consider black-and-white images, $X = \{0,1\}^N$, and two subsets:

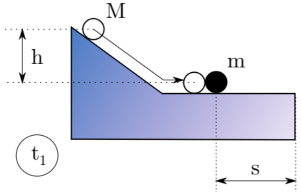


- **possible states of a gas starting from a given initial state**
  - very chaotic images, $|C_a^{M,epsilon}|$ very large
  - only **few relevant** quantities (thermodynamics: V, E, N)
  - all other are irrelevant
- **black square on white background**
  - very ordered images, $|C_a^{M,epsilon}|$ very small
  - “square”: collection of **lot of laws** (relevant features)
  - only **three irrelevant** quantities (x, y, a)

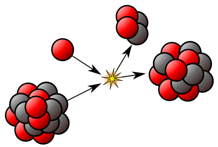


Number of relevant features

Spectrum in number of relevant coordinates



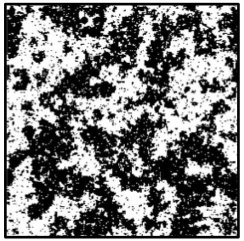
point mechanics
~ 5 relevant



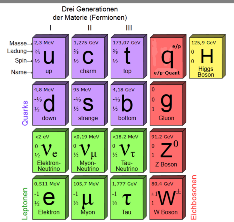
nuclear physics
20-? relevant

chemistry, biology
~ 100-? relevant

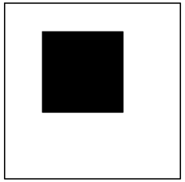
natural environment
? relevant ? irrelevant



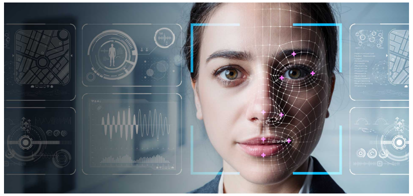
Ising model
3 relevant



Standard Model
21 relevant
(symmetries!)



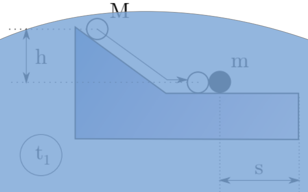
geometric images
~ 10-100 irrelevant



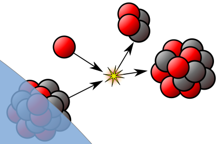
face recognitions
~ 30000 irrelevant

Number of relevant coordinates

Spectrum in number of relevant coordinates



point mechanics
~ 5 relevant

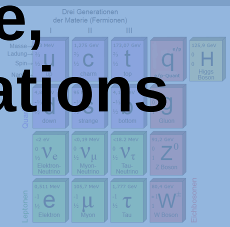


nuclear physics
20-? relevant

chemistry, biology
~ 100-? relevant

natural environment
? relevant ? irrelevant

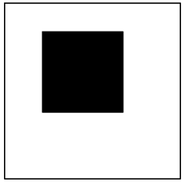
Few relevant quantities, identifyable, exact laws, relations



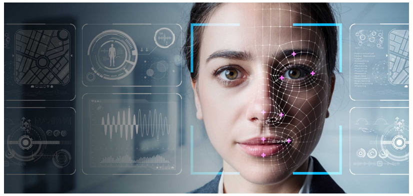
SCIENCE

Ising model
3 relevant

Standard Model
21 relevant
(symmetries!)



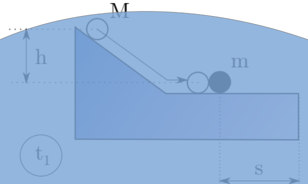
geometric images
~ 10-100 irrelevant

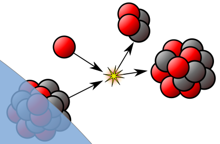


face recognitions
~ 30000 irrelevant

Number of relevant coordinates

Spectrum in number of relevant coordinates





point mechanics
~ 5 relevant

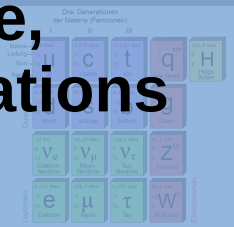
nuclear physics
20-? relevant

chemistry, biology
~ 100-? relevant

nature, climate, ...
? relevant ? irrelevant

Few relevant quantities,
identifyable,
exact laws, relations

Lots of relevant quant.
unidentifyable,
similar role,
no exact description



SCIENCE

(ARTIFICIAL)
INTELLIGENCE

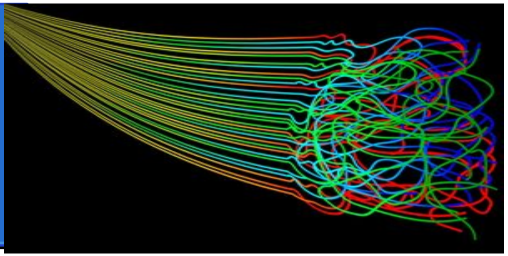
Ising model
3 relevant

Standard Model
21 relevant
(symmetries!)

geometric images
~ 10-100 irrelevant

face recognitions
~ 30000 irrelevant

Entropy of the intelligence



Intelligence or understanding is the choice of correct representation.

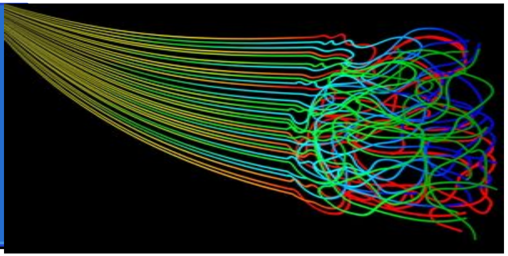
Is there a universal measure to decide, how good a given representation is?

entropy of a representation with respect to a subset

- **Shannon entropy:** $$S_{SH}=\sum_{\sigma\in B^N} p_C(\xi=\sigma)\log_2 p_C(\xi=\sigma)=\log_2|C|$$
  - independent of the representation
  - yields the true information content of the set (i.e. the number of necessary bits)
- **representation entropy:** ξ coordination implies $p_C(\xi_i=\sigma_i)$ **bitwise distribution**

$$S_{repr}=\sum_{i=1}^{N}\Big[\sum_{\sigma\in 0,1} p_C(\xi_i=\sigma_i)\log_2 p_C(\xi_i=\sigma)\Big]$$

Entropy of the intelligence



Representation entropy

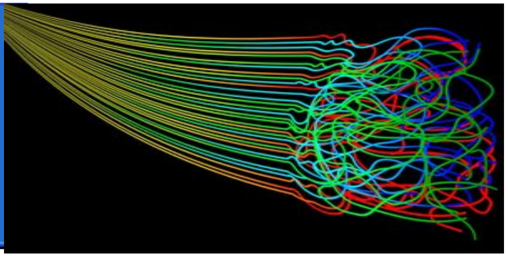
$$S_{repr} = \sum_{i=1}^{N} \left[\sum_{\sigma \in 0,1} p_C(\xi_i = \sigma_i) \log_2 p_C(\xi_i = \sigma) \right]$$

Mathematical properties

- $S_{repr} \geq S_{SH}$, equality iff the coordination is independent
- minimality of S_{repr} implies independence, and the least descriptive coordinates
- for our “cat” images $S_{SH} = \log_2 4 = 2$
  - in the original representation C={0001,0110,1010,1011} ⇨ $S_{repr} = 3.62$
  - in the proper representation C={0000,0001,0010,0011} ⇨ $S_{repr} = 2$

representation entropy is a general unsupervised loss function:
in a general learning process, by minimizing the representation entropy, we get closer to the learning of the proper representation

Entropy of the intelligence



Practical improvements

$$Loss = S_{repr} + \lambda\,\alpha + \mu\,\beta$$

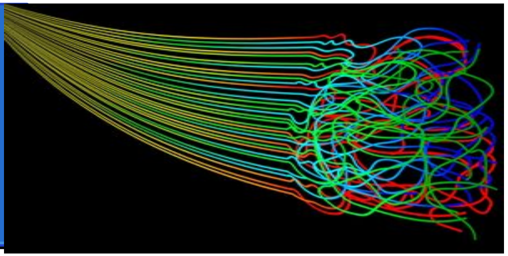
- α is type one error (false negative): $\xi(x) = \xi(C)$ for selective coordinates, but $x \notin C$
- β is type two error (false positive): $\xi(x) \neq \xi(C)$ for selective coordinates, but $x \in C$

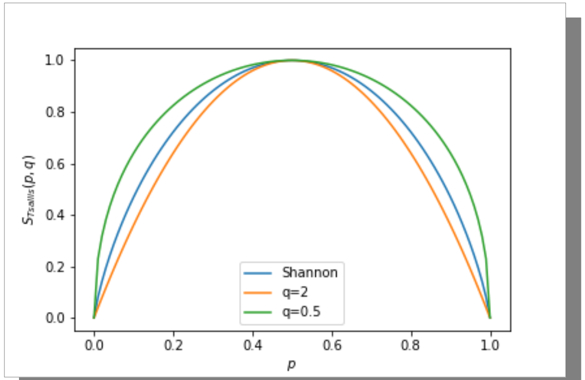
$$\alpha = P_X(\{x \in X \mid \xi_{rel}(x) = \xi_{rel}(C), x \notin C\})$$

$$\beta = P_X(\{x \in X \mid \xi_{rel}(x) \neq \xi_{rel}(C), x \in C\})$$

- for perfect coordination $\alpha = \beta = 0$
- in practical applications with correct choice of coefficients we can improve convergence

Entropy of the intelligence





Representation entropy, generalized

$$S_{repr} = \sum_{i=1}^{N} S(p_C(\xi_i))$$

we can use generalized entropy formulae

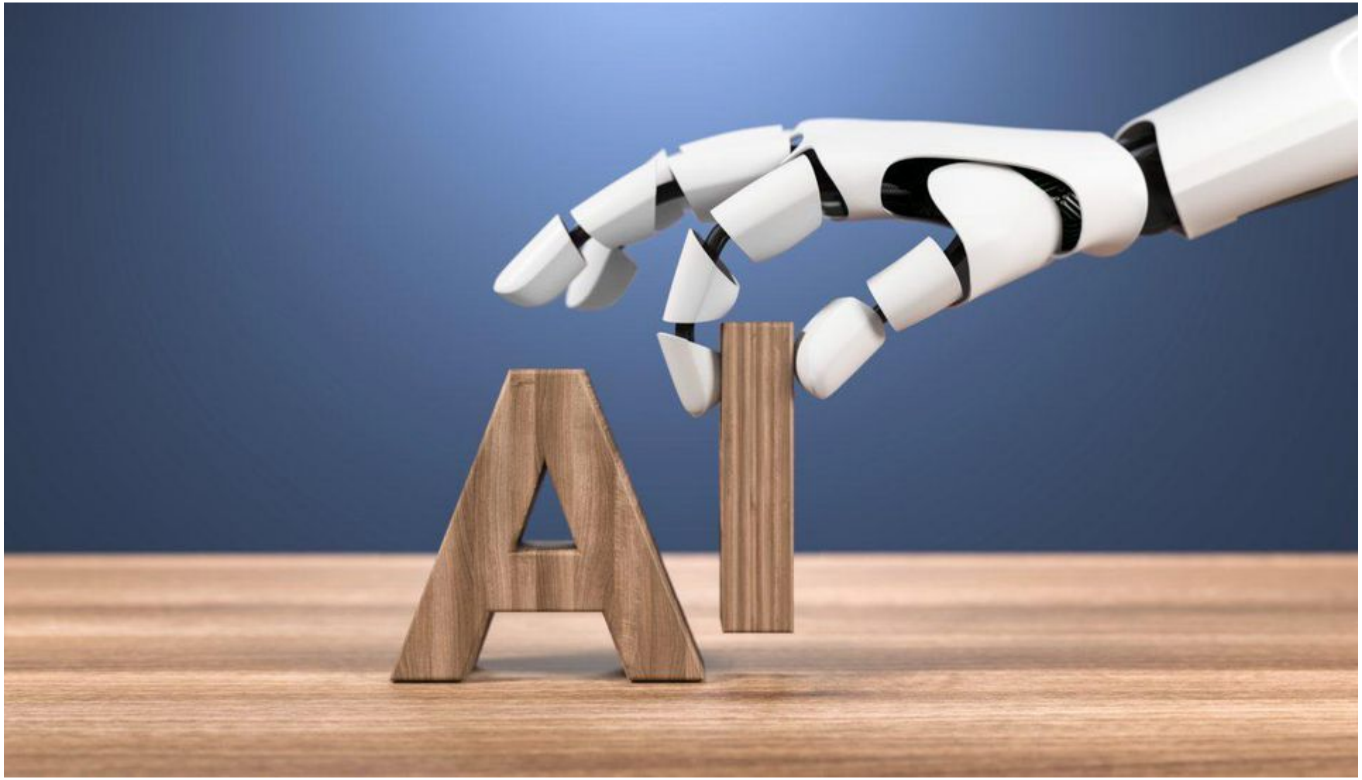
- Tsallis form: $S(p(\xi)) = \frac{\alpha}{q-1}\left(1 - \sum_{\sigma \in B} p(\xi = \sigma)^q\right)$
- because of the normalization $S_{repr} = S_{SH}$ for independent coordination
- main difference is the slope at p=0,1: how preferred are the selective coordinates
- for $q > 1$ if there is enough memory and |C| is large enough, it is worth to memorize the elements one-by-one, e.g.: C={00000, 00001, 00010, 00100, 01000, 10000}

Conclusions

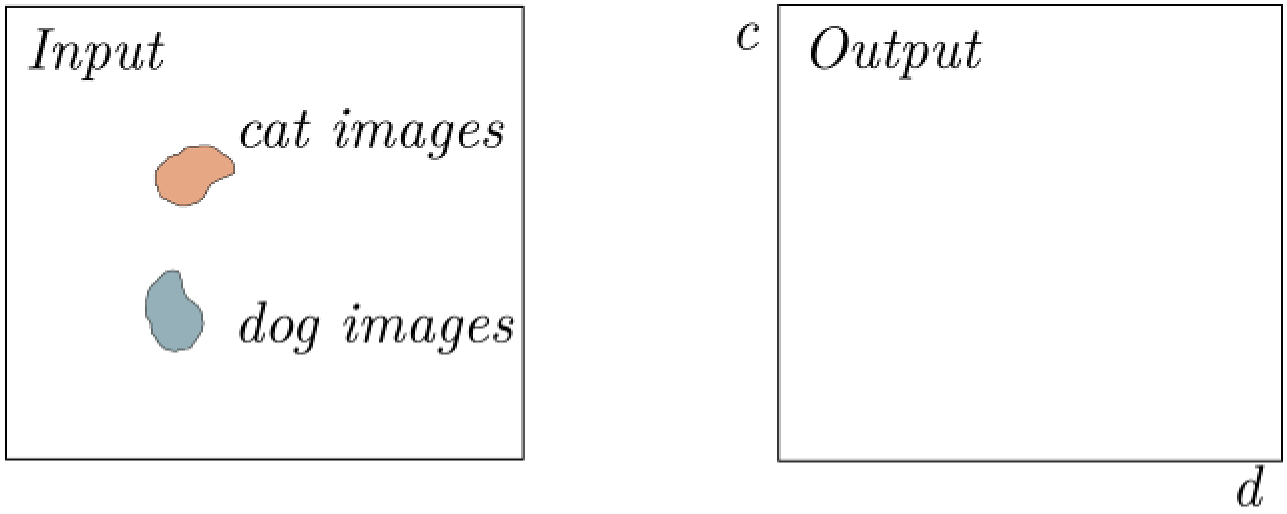
understanding ≡ best representation of data

- independent features (coordinates) over a set C: either selective or descriptive
- selective/relevant features: constant over C, good for classification
- descriptive/irrelevant features: variable over C, good for compression
- *number of relevant coordinates* can vary vastly
  - few (Ising model): adequate for scientific modeling
  - lot (natural images): adequate for (artificial) intelligence modeling
- representation entropy: universal unsupervised loss function, by minimizing it we improve understanding.

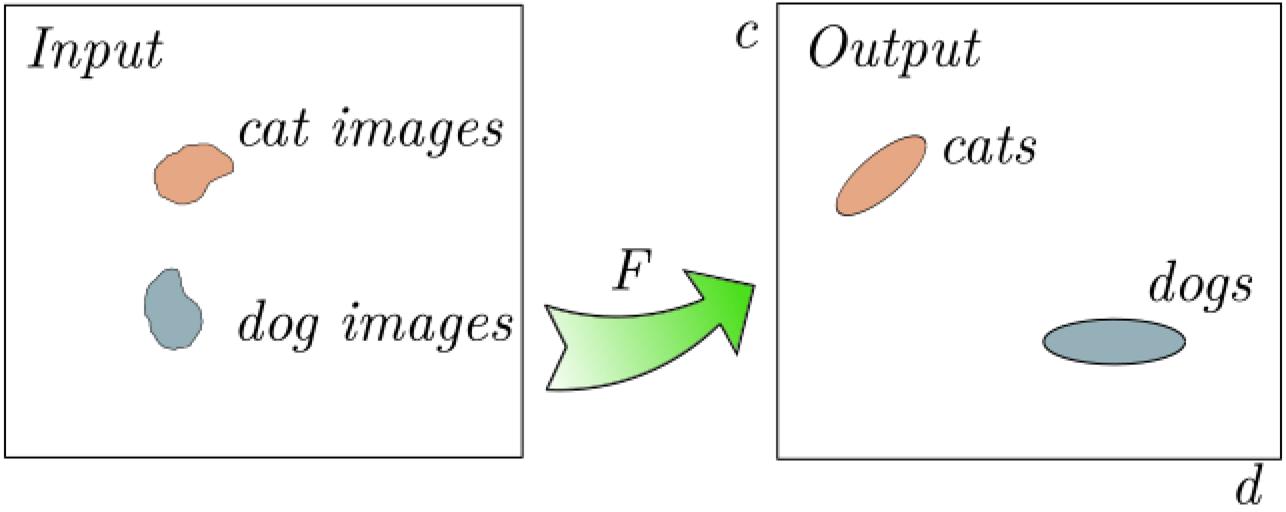
The end



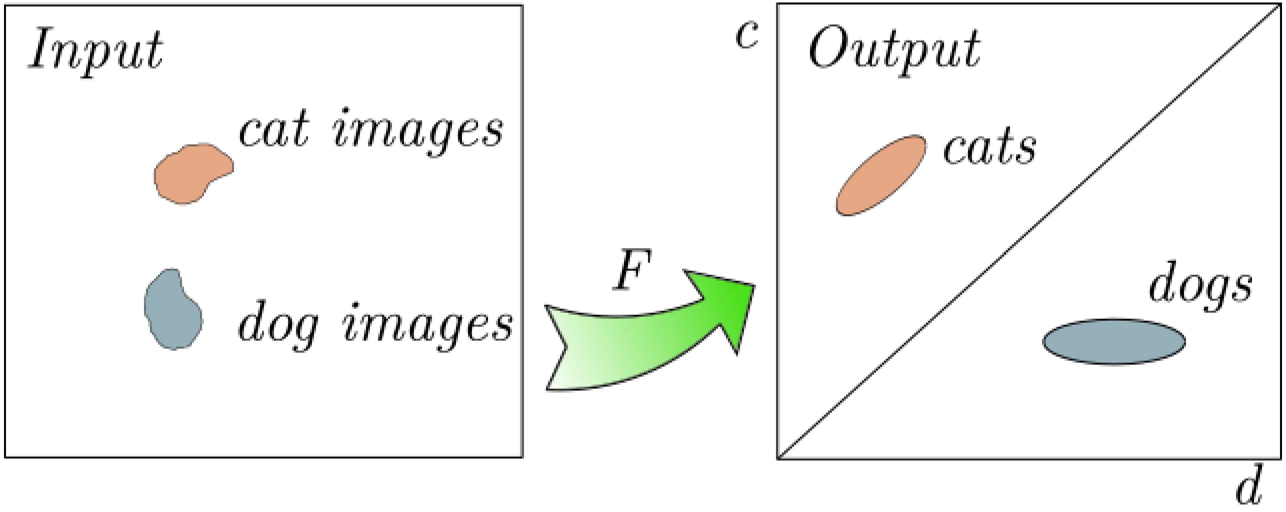
Interpretation of AI



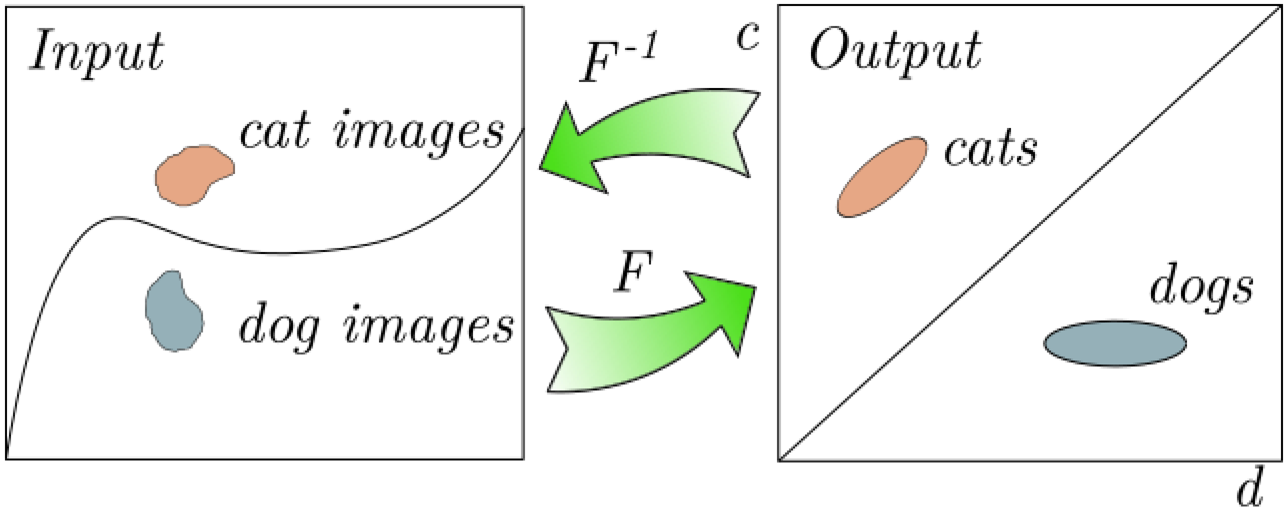
Interpretation of AI



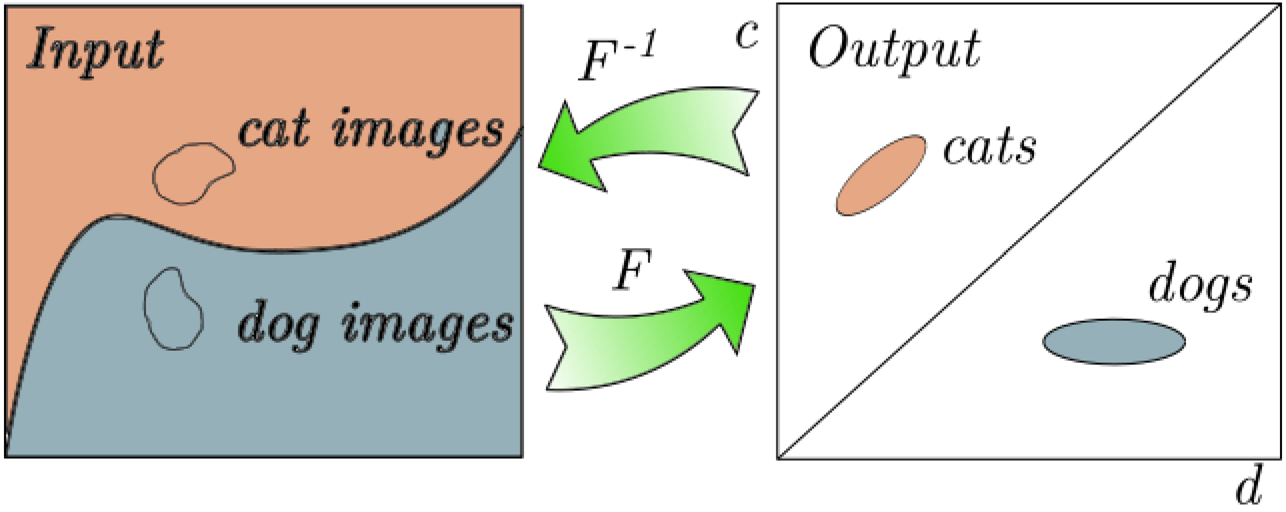
Interpretation of AI



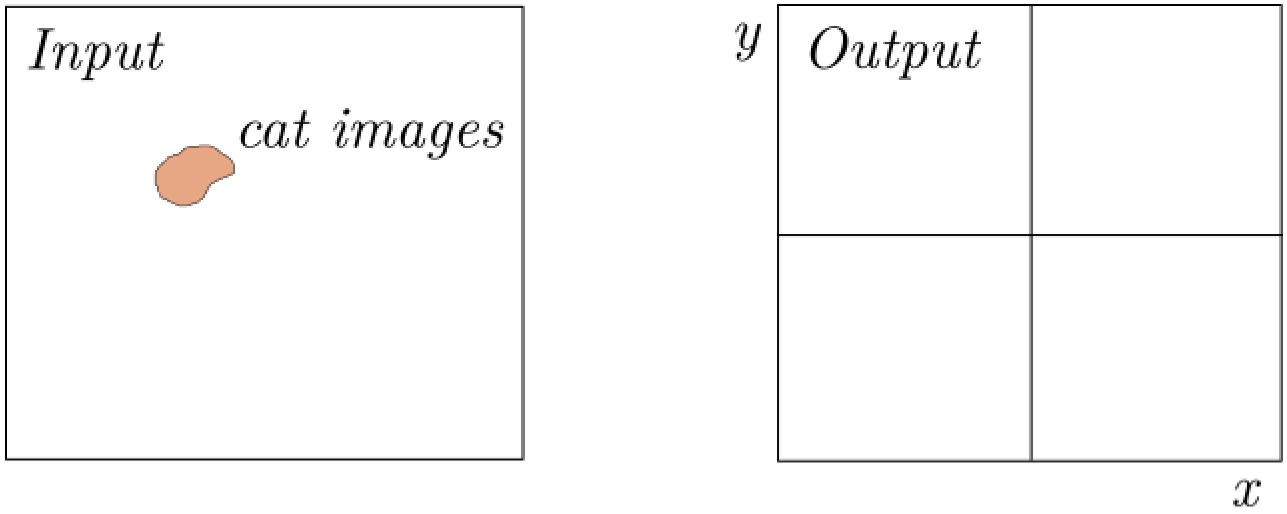
Interpretation of AI



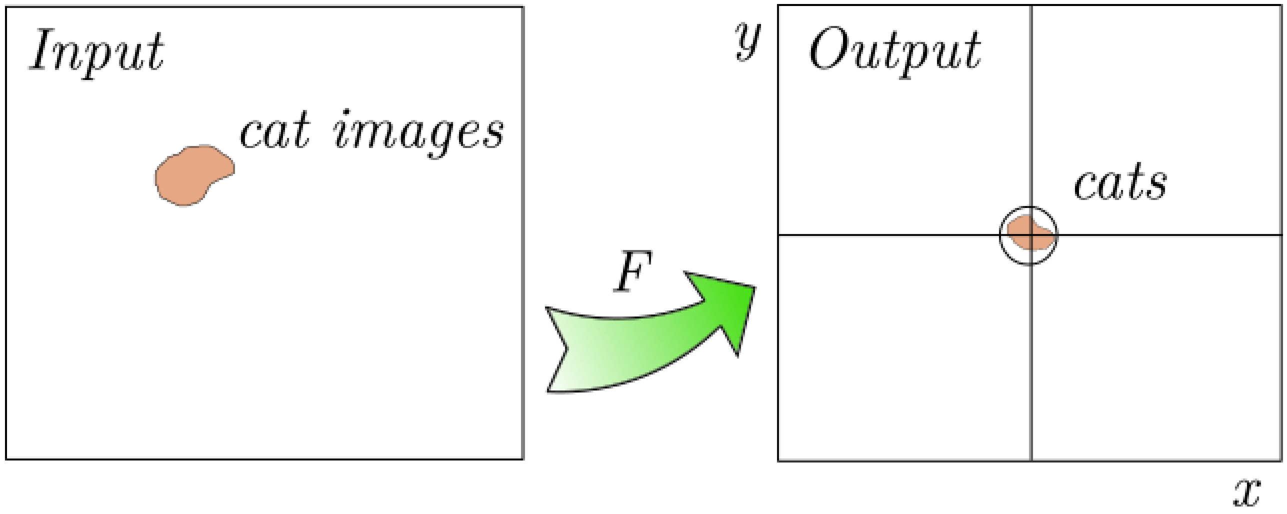
Interpretation of AI



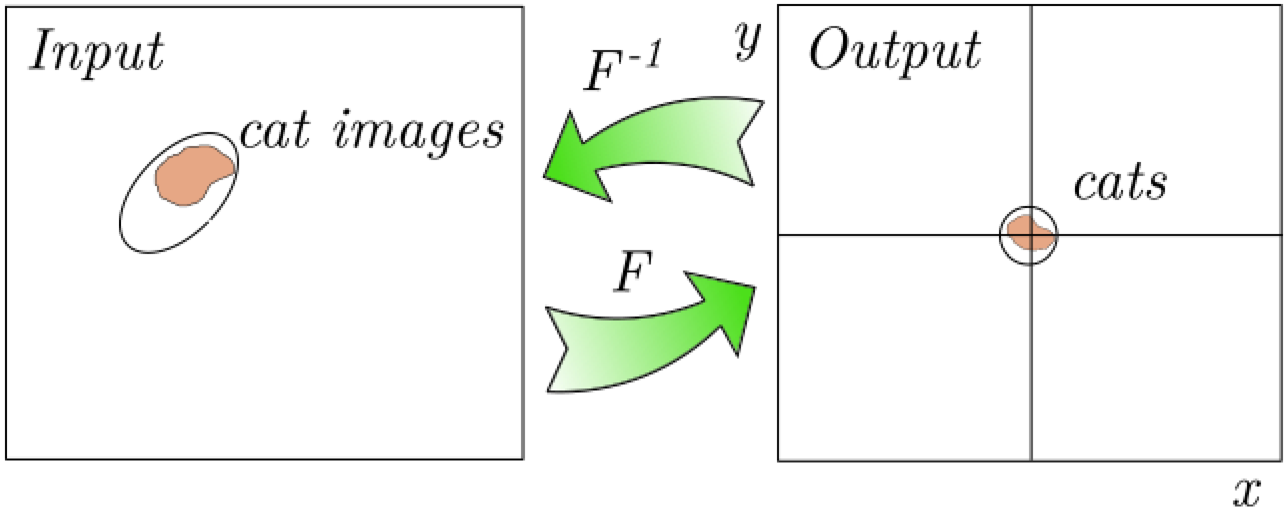
Interpretation of AI



Interpretation of AI



Interpretation of AI



Interpretation of AI