Lecture 02b:

What is Entropy?

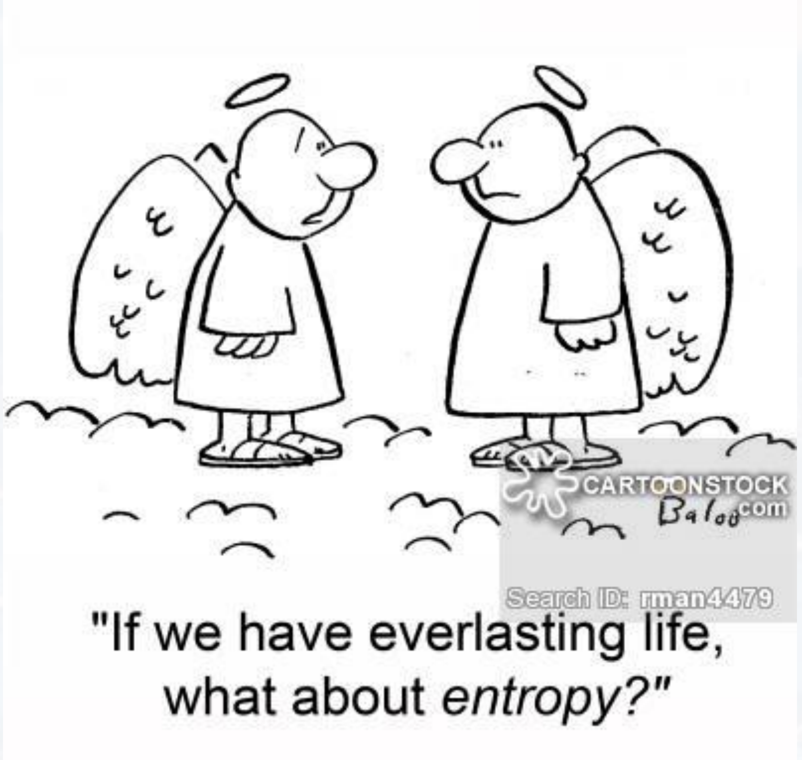
Markus Hohle

University California, Berkeley

Machine Learning Algorithms

MSSE 277B, 3 Units

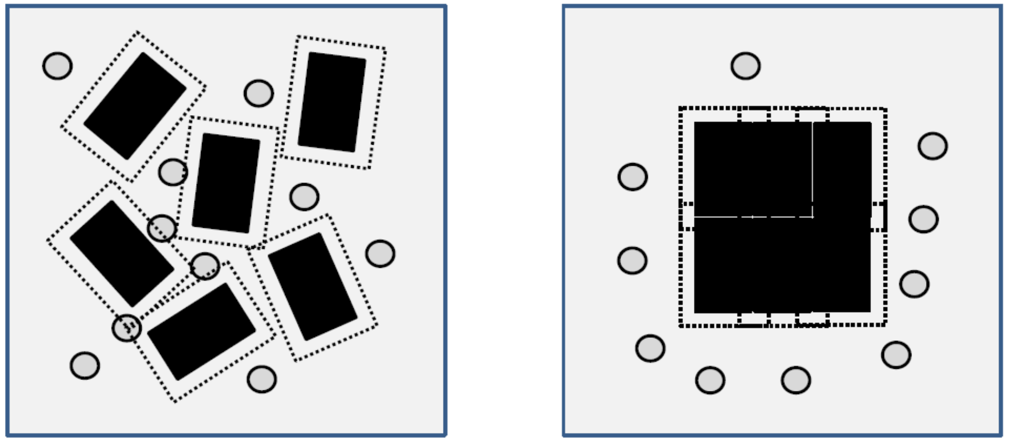
Spring 2025



"If we have everlasting life, what about *entropy*?"

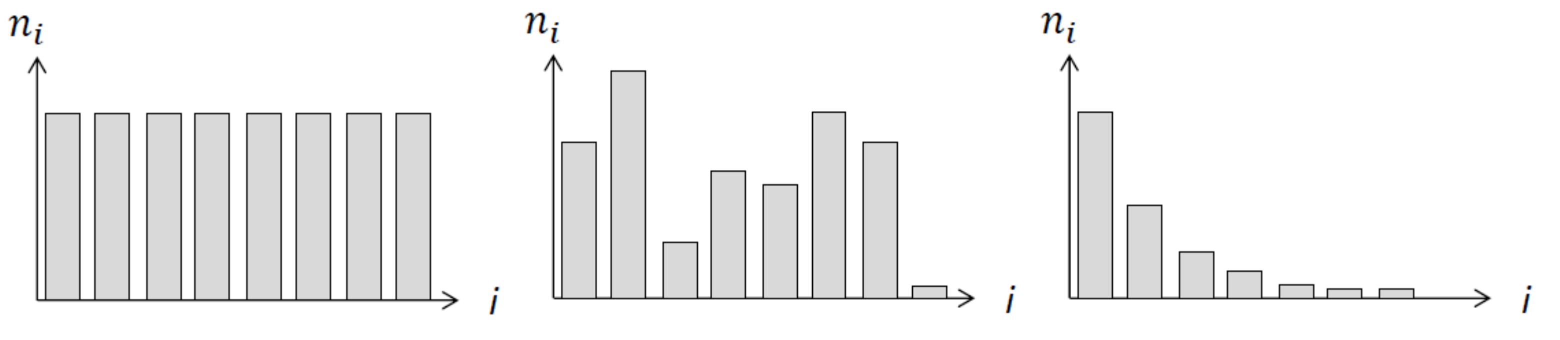
What do you think: in which image is entropy higher?

first of all:
Entropy is **not** a measure of disorder!



low entropy high entropy

i: states
n_i: number of particles in state i



highest **possible** entropy low entropy high entropy

Often people explain entropy with an ordered vs messy office…

…and then say, that entropy (disorder) grows with time (in closed systems).

first of all:

Entropy is **not** a measure of disorder!

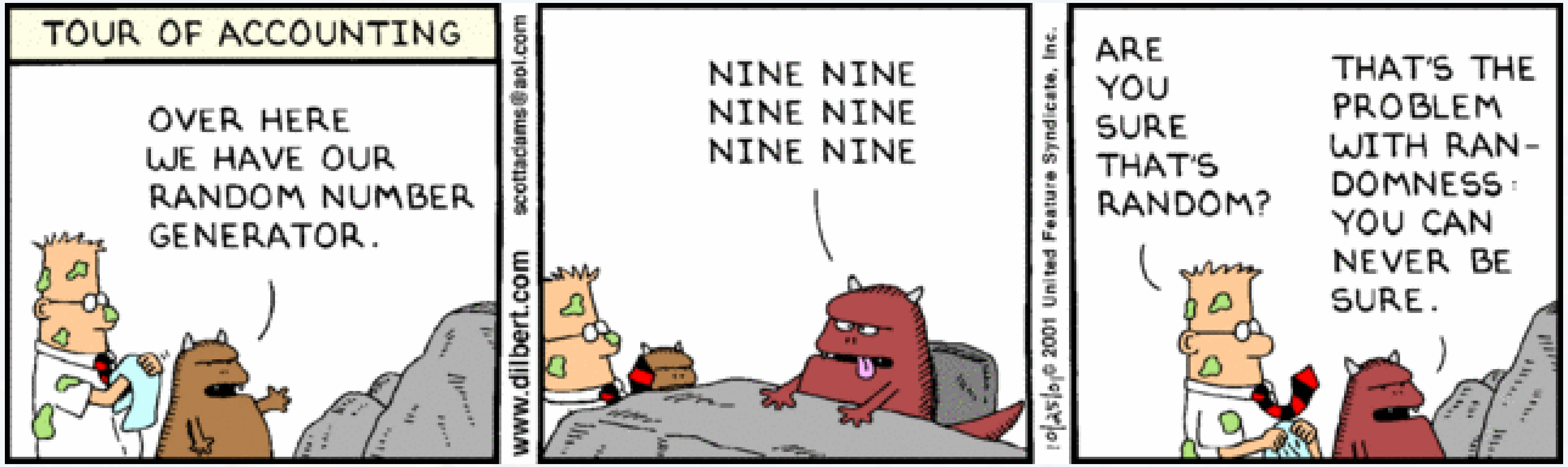
- But how is it possible, that an office can do that, just by itself?
- What if my office just *looks* messy, but I can still pull any file you are asking me for?

order/disorder is not a physical quantity!

Those examples have nothing to do with entropy conceptionally!

Machine Learning Algorithms: Entropy

actually, the idea of entropy is more like that:

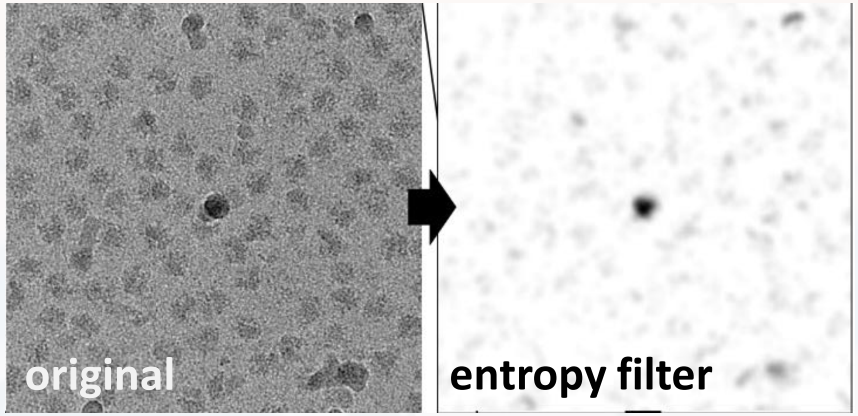


Machine Learning Algorithms: Entropy

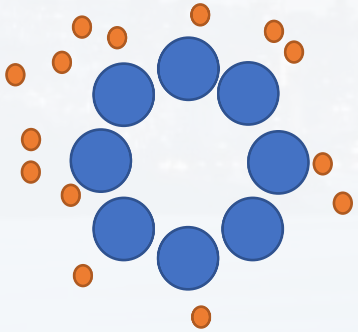
Entropy:

data analysis:
- image processing
- noise reduction
- feature detection

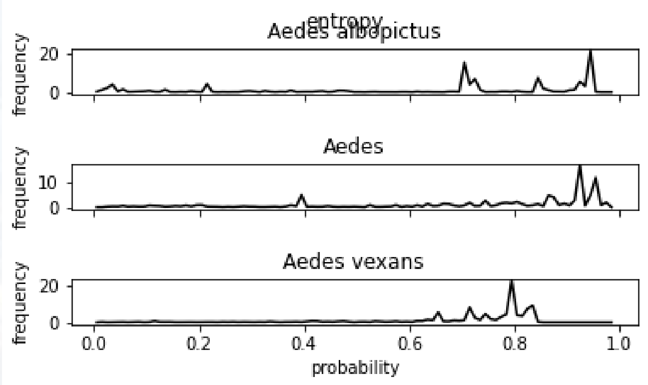
Cryo-EM image of ribosomes



biophysics:
- molecular driving forces
- formation of macromolecules
- “ordering forces”



AI:
- optimization
- cross entropy



Entropy: **statistics/information theory:** - maximum entropy, given constrains

Distribution name	Probability density / mass function	Maximum Entropy constraint	Support
Uniform (discrete)	$f(k)=\frac{1}{b-a+1}$	None	$\{a,a+1,\ldots,b-1,b\}$
Uniform (continuous)	$f(x)=\frac{1}{b-a}$	None	$[a,b]$
Bernoulli	$f(k)=p^k(1-p)^{1-k}$	$\mathbb{E}[\,K\,]=p$	$\{0,1\}$
Geometric	$f(k)=(1-p)^{k-1}\,p$	$\mathbb{E}[\,K\,]=\frac{1}{p}$	$\mathbb{N}\setminus\{0\}=\{1,2,3,\ldots\}$
Exponential	$f(x)=\lambda\exp(-\lambda x)$	$\mathbb{E}[\,X\,]=\frac{1}{\lambda}$	$[0,\infty)$
Laplace	$f(x)=\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	$\mathbb{E}[\,\lvert X-\mu\rvert\,]=b$	$(-\infty,\infty)$
Asymmetric Laplace	$f(x)=\frac{\lambda\,\exp\left(-(x-m)\,\lambda\,s\,\kappa^s\right)}{(\kappa+\frac{1}{\kappa})}$ where $s\equiv\operatorname{sgn}(x-m)$	$\mathbb{E}[\,(X-m)\,s\,\kappa^s\,]=\frac{1}{\lambda}$	$(-\infty,\infty)$
Pareto	$f(x)=\frac{\alpha\,x_m^\alpha}{x^{\alpha+1}}$	$\mathbb{E}[\,\ln X\,]=\frac{1}{\alpha}+\ln(x_m)$	$[x_m,\infty)$
Normal	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{E}[\,X\,]=\mu$, $\mathbb{E}[\,X^2\,]=\sigma^2+\mu^2$	$(-\infty,\infty)$
Gamma	$f(x)=\frac{x^{k-1}\exp\left(-\frac{x}{\theta}\right)}{\theta^k\,\Gamma(k)}$	$\mathbb{E}[\,X\,]=k\,\theta$, $\mathbb{E}[\,\ln X\,]=\psi(k)+\ln\theta$	$[0,\infty)$

What is entropy, really?

N:	number of dice
n_i:	number of dice exposing a certain number *i* (= having a certain state *i*)
I:	number of states a die can have

What is the probability *P* to observe the *system* in a certain state?

What is the probability p_i to observe *a die* in a certain state?

$$\Omega = \frac{N!}{n_1!\,n_2!\ldots n_I!}$$

assumption: all *i* are equally likely: $P = 1/\Omega$

Machine Learning Algorithms:

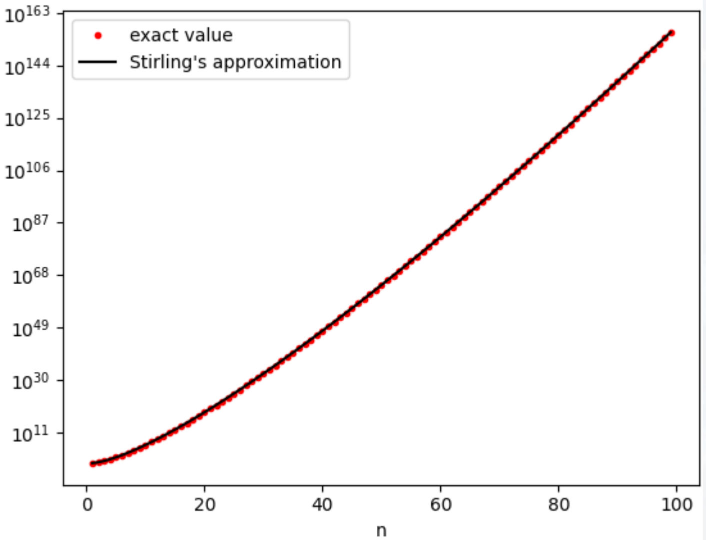
***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**

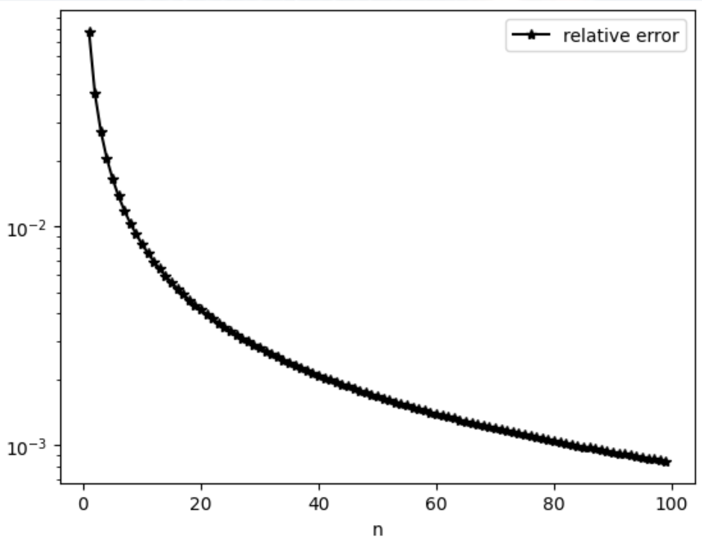
$$\Omega = \frac{N!}{n_1!\, n_2! \ldots n_I!}$$

is large, even for small systems!

Stirling's approximation

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$





N:	number of dice
$\boldsymbol{n_i}$:	number of dice exposing a certain number ***i*** (= having a certain state ***i***)
I:	number of states a die can have

$$\Omega = \frac{N!}{n_1!\, n_2! \ldots n_I!}$$ is large, even for small systems!

for large n_i:

$$\Omega = \frac{N!}{n_1!\, n_2! \ldots n_I!} \approx \frac{N^N}{{n_1}^{n_1} {n_2}^{n_2} \ldots {n_I}^{n_I}}$$

$$p_i \approx \frac{n_i}{N}$$

$$\approx \frac{1}{{p_1}^{n_1} {p_2}^{n_2} \ldots {p_I}^{n_I}}$$

$$ln\Omega = -\sum_i^I n_i \, lnp_i \qquad \frac{\ln\Omega}{N} = -\sum_i^I p_i \, lnp_i \qquad \boxed{S = -\sum_i^I p_i \, lnp_i}$$

entropy per particle

Machine Learning Algorithms: Entropy

$$S = -\sum_{i}^{I} p_i \, lnp_i$$

N:	**number of dice**
n_i:	**number of dice exposing a certain number *i*** **(= having a certain state *i*)**
I:	**number of states a die can have**
p_i:	n_i**/N**

assumption: all *i* are equally likely: $P = 1/\Omega$

<u>subsets of Ω:</u>

- sum *M* of all numbers on the dice
- dice can only be distinguished by their state

N = 2:

min(M) = 1 + 1 = 2

max(M) = 6 + 6 = 12

most likely ***M = 7*** (or *2 x mean(I)*), because there are **six** possibilities to obtain it:

1 + 6; 1 + 6; 2 + 5; 5 + 2; 3 + 4; 4 + 3

N:

min(M) = N

max(M) = I M

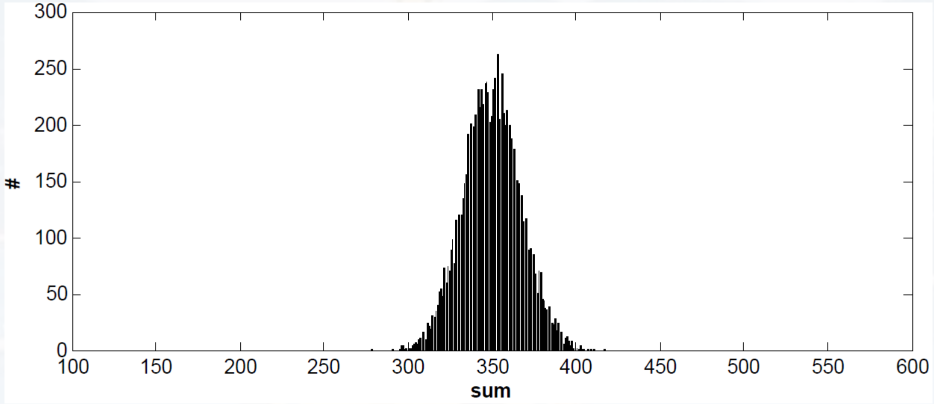
most likely ***M = N x mean(I)***

Machine Learning Algorithms:

$$S = -\sum_{i}^{I} p_i \, lnp_i$$

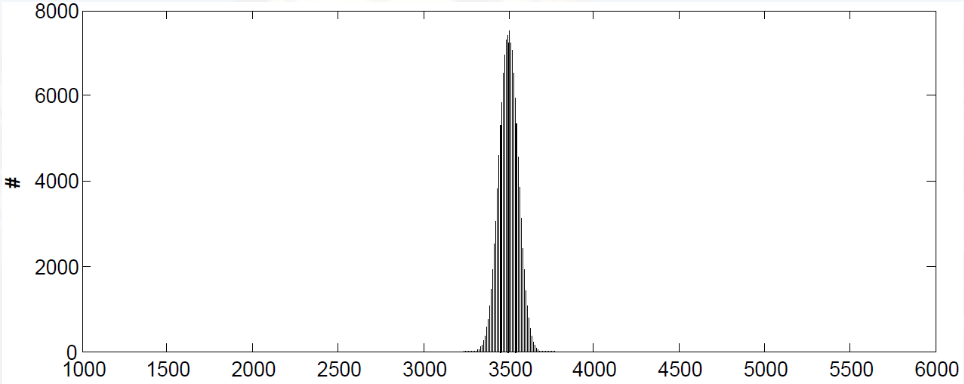
N:	**number of dice**
$\boldsymbol{n_i}$:	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
$\boldsymbol{p_i}$:	$\boldsymbol{n_i}$**/N**

assumption: all *i* are equally likely:



N = 100

- some subsets of Ω, hence some states of the system are **way more likely** than other states
- becomes **more extreme for large *N***

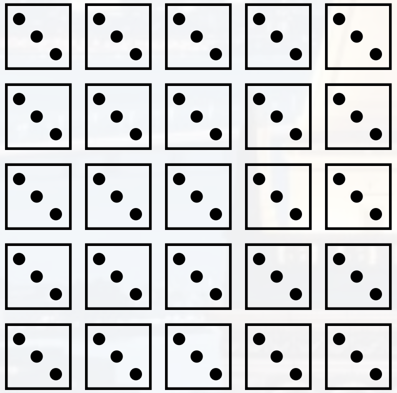


N = 1000

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
p_i:	**n_i/N**

<u>we can also see this as dynamical process:</u>



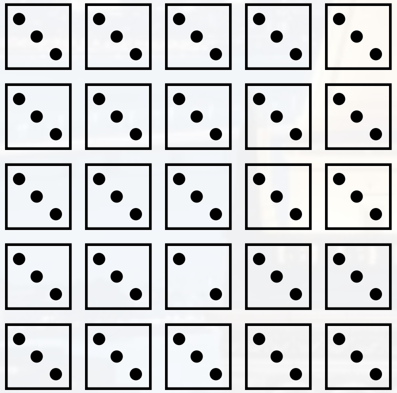
t = 0 : all dice have the same state

t → t + dt : one, randomly picked die, changes its state

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
p_i:	**n_i/N**

we can also see this as dynamical process:



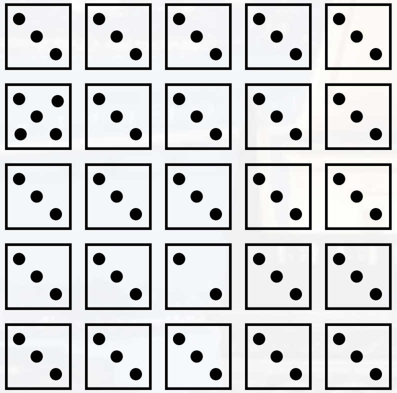
t = 0 : all dice have the same state

t → t + dt : one, randomly picked die, changes its state

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i*** **(= having a certain state *i*)**
I:	**number of states a die can have**
p_i:	n_i**/N**

<u>we can also see this as dynamical process:</u>



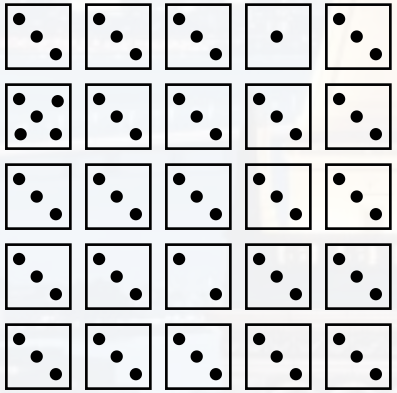
t = 0 : all dice have the same state

t → t + dt : one, randomly picked die, changes its state

$$S = -\sum_{i}^{I} p_i \, lnp_i$$

***N*:**	**number of dice**
n_i**:**	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
p_i**:**	n_i**/N**

<u>we can also see this as dynamical process:</u>



t = 0 : all dice have the same state

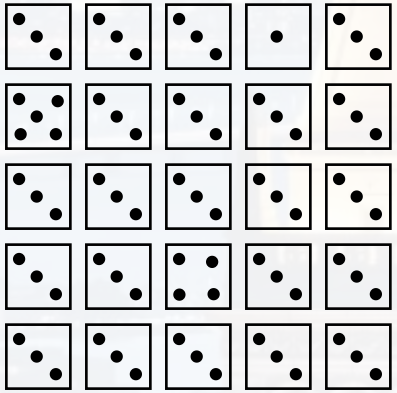
t → t + dt : one, randomly picked die, changes its state

Machine Learning Algorithms: Entropy

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

***N*:**	**number of dice**
n_i**:**	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
p_i**:**	n_i**/N**

<u>we can also see this as dynamical process:</u>



t = 0 : all dice have the same state

t → t + dt : one, randomly picked die, changes its state

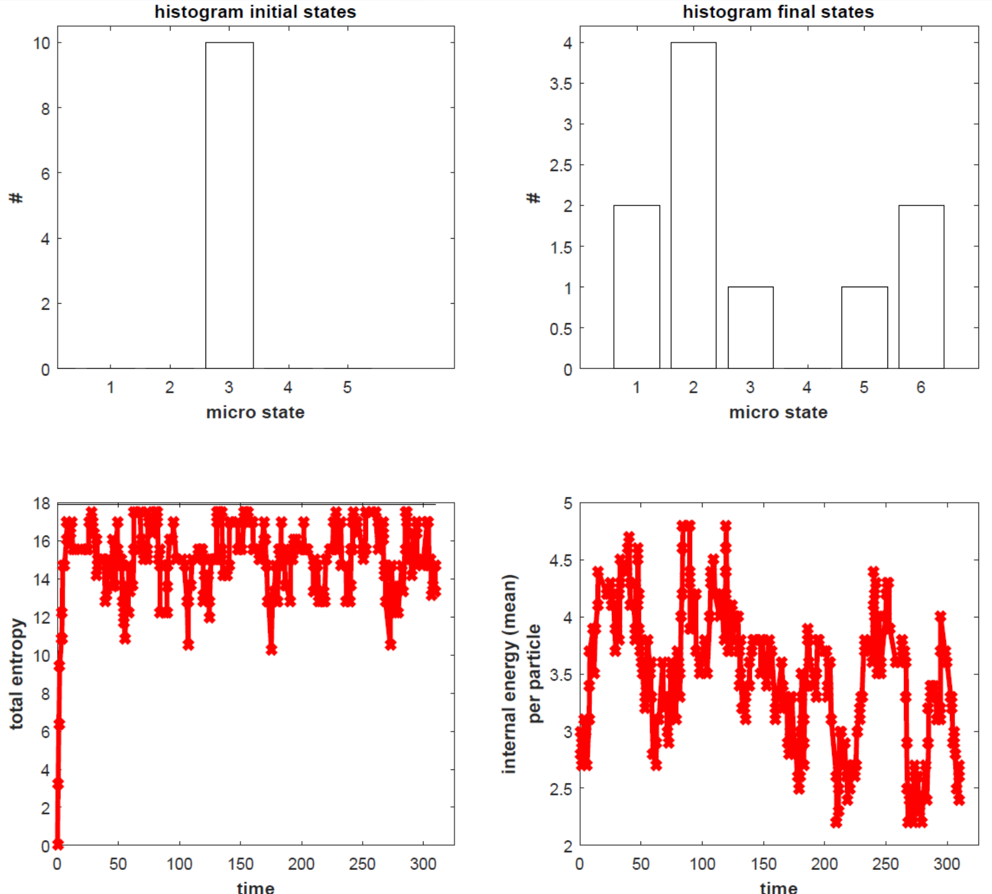
Machine Learning Algorithms:

Entropy

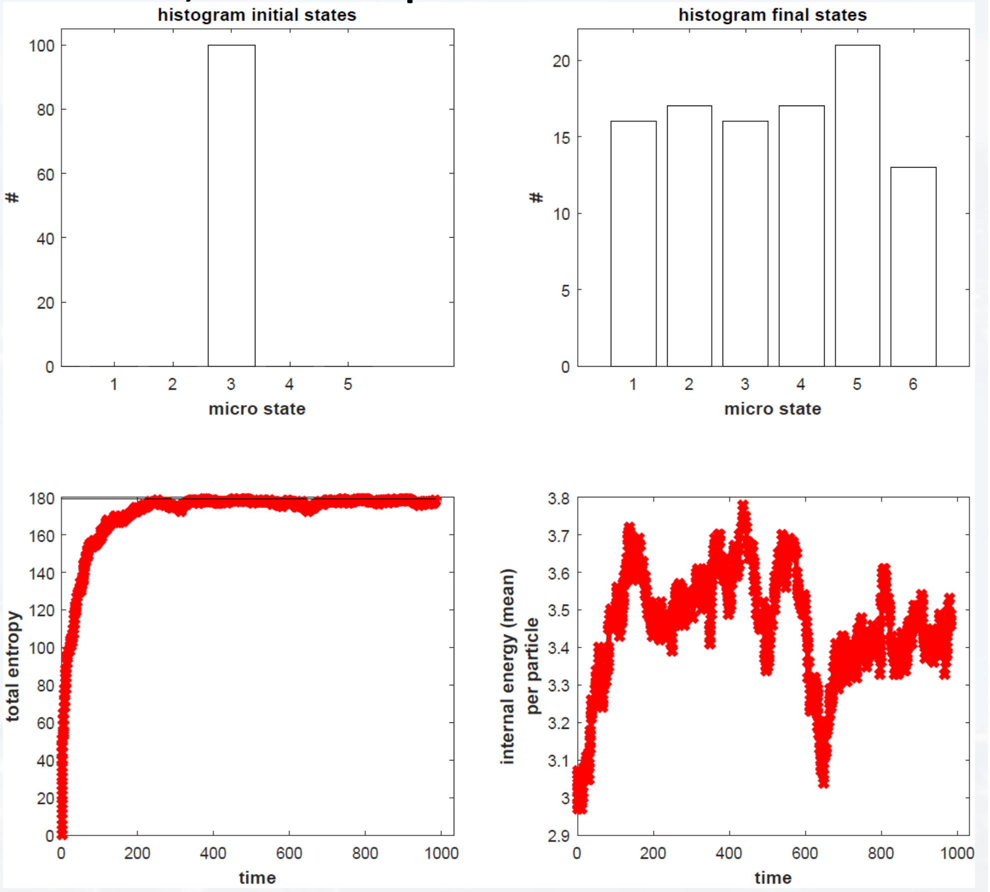
$$S = -\sum_{i}^{I} p_i \, lnp_i$$

***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i***
	(= having a certain state *i*)
I:	**number of states a die can have**
p_i:	**n_i/N**

10 dice
300 timesteps



100 dice 1,000 timesteps

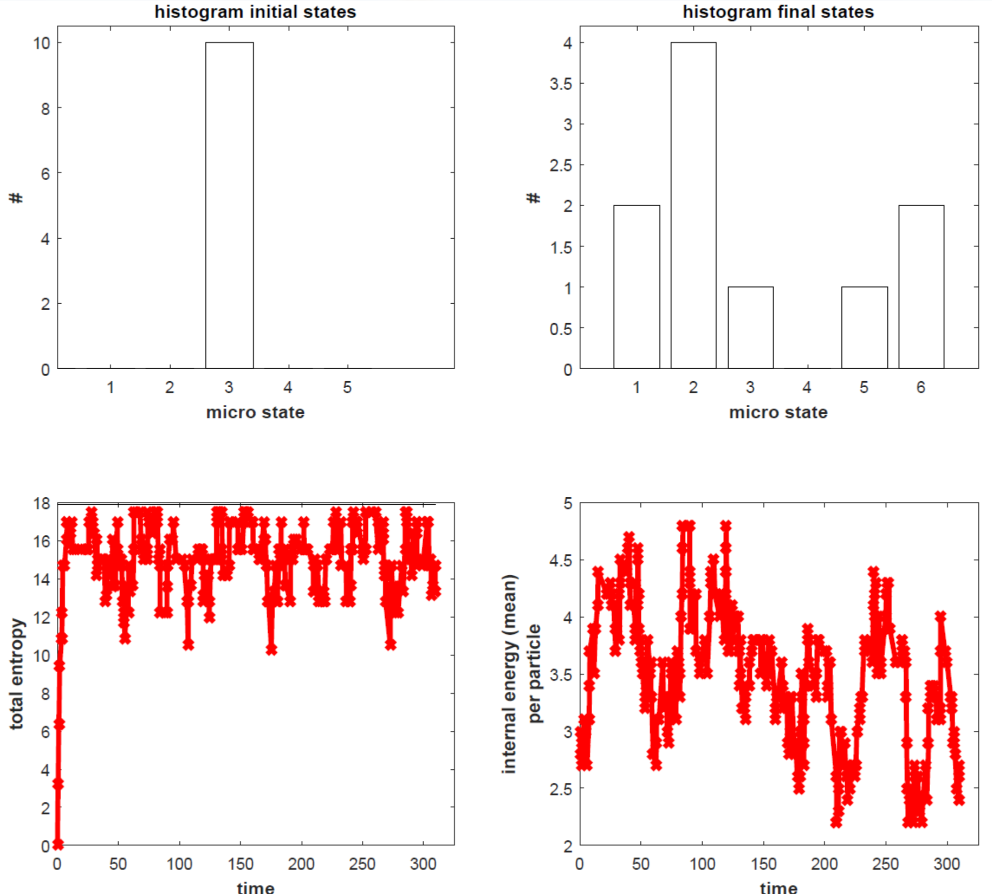


Machine Learning Algorithms: Entropy

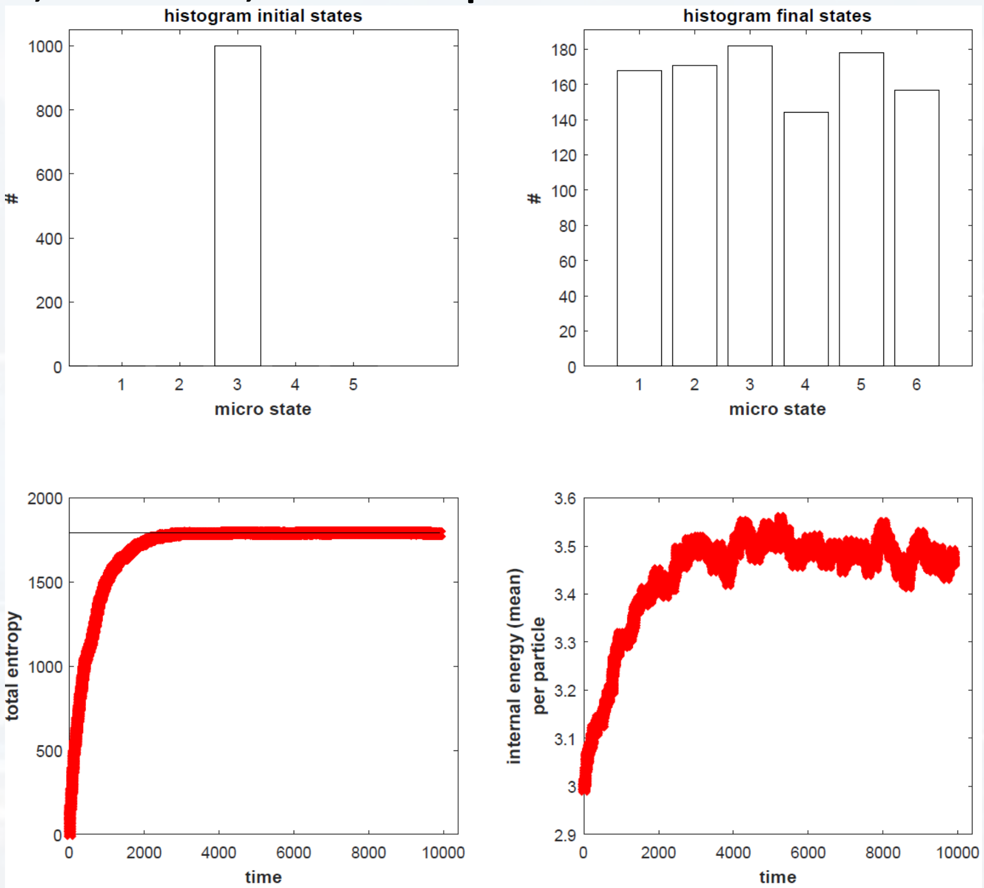
$$S = -\sum_{i}^{I} p_i \, ln p_i$$

***N*:**	**number of dice**
n_i:	**number of dice exposing a certain number *i* (= having a certain state *i*)**
I:	**number of states a die can have**
p_i:	**n_i/N**

10 dice
300 timesteps

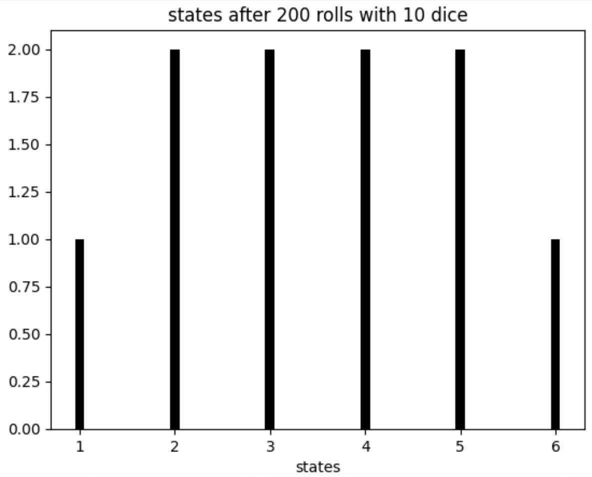


1,000 dice 10,000 timesteps

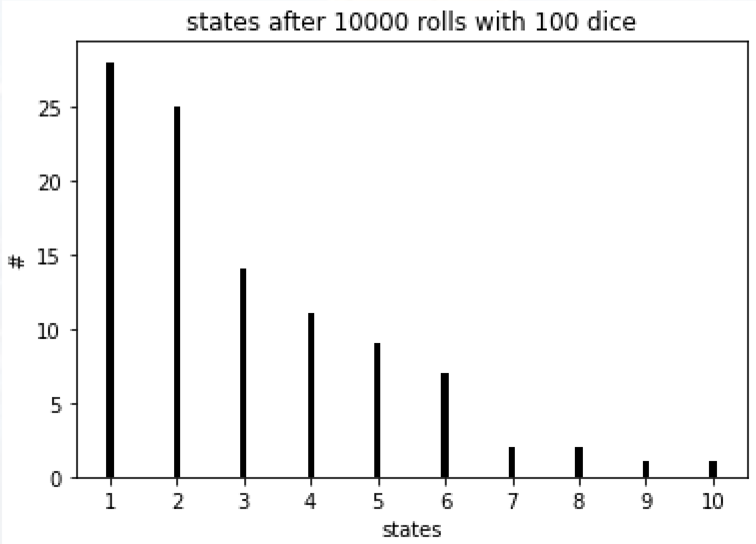


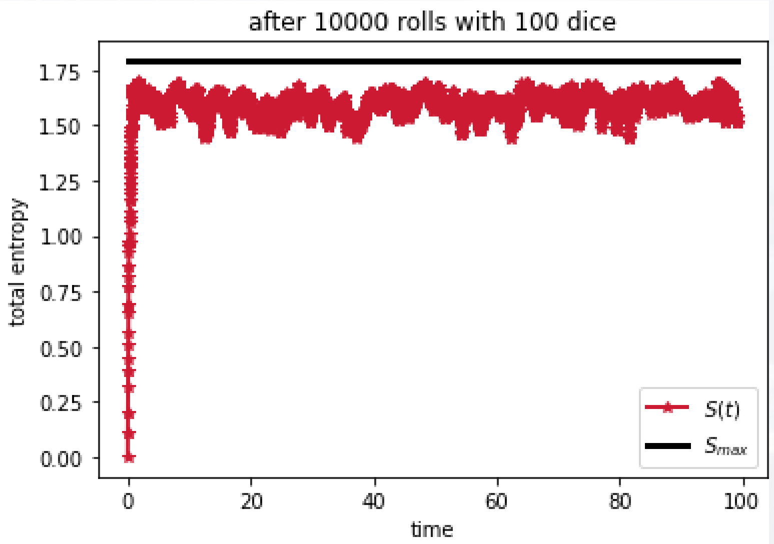
Machine Learning Algorithms:

check out `random_machine.ipynb`



constrain: *M* is conserved, *I* is free, but >0





conclusions:

- entropy increases with time **(in a closed system)** because it is the **most likely state**

- the **larger** the system, the more **deterministic** it looks

- for **small systems**: entropy can fluctuate in both ways and **does not increase**!
  (Stirling's approximation)

- **large systems**: (thermodynamic) **arrow of time**
  (question: what if we are at S_{max} already)

- **small systems**: **symmetry** in time!

- even if *i* are not equally likely (constrains, some states are more accessible)
  → different weights, **but same principle**

- **uniform distribution** has **highest entropy** (do the math :))

Entropy is a mathematically precise measure of information!

- entropy high → information low
- entropy low → information high

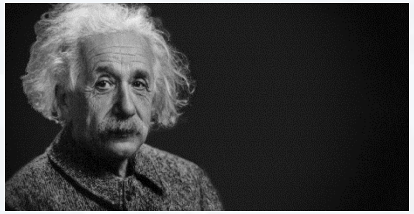
$$S = -\sum_{i}^{I} p_i \, ln p_i$$

fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Hi Eini: quess the number I rolled!

I have no idea, so all numbers are equally likely if it's a fair die.



Machine Learning Algorithms: Entropy

Entropy is a mathematically precise measure of information!

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

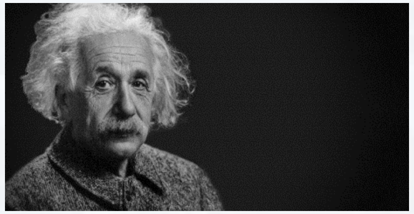
- entropy high → information low
- entropy low → information high

fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Hi Eini: quess the number I rolled!

$$\boldsymbol{S = -\frac{1}{6} ln\left(\frac{1}{6}\right) * 6 = 1.79 \ldots}$$



Entropy is a mathematically precise measure of information!

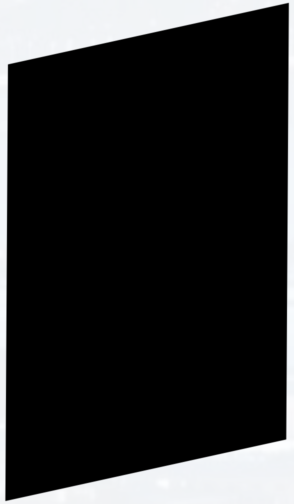
- entropy high → information low
- entropy low → information high

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

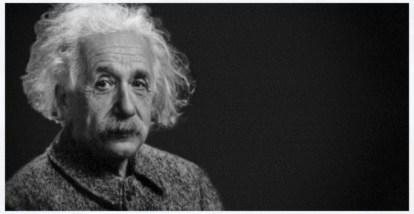
<u>**fair** die:</u> events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Ok, I give you a hint: it is not a six.



Alright, so then $p_6 = 0$ and all the other $p_5 = 1/5$.



Machine Learning Algorithms:

Entropy is a mathematically precise measure of information!

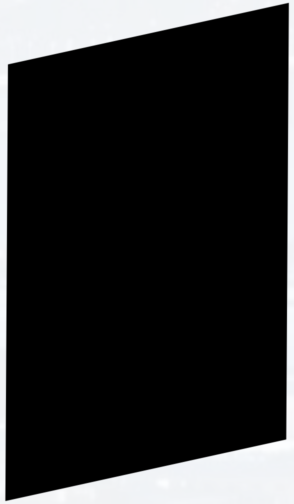
- entropy high → information low
- entropy low → information high

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

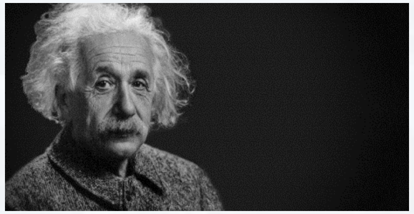
fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Ok, I give you a hint: it is not a six.



Hence, $\boldsymbol{S = -\frac{1}{5} ln\left(\frac{1}{5}\right) * 5 = 1.61 \ldots}$



Machine Learning Algorithms: Entropy

Entropy is a mathematically precise measure of information!

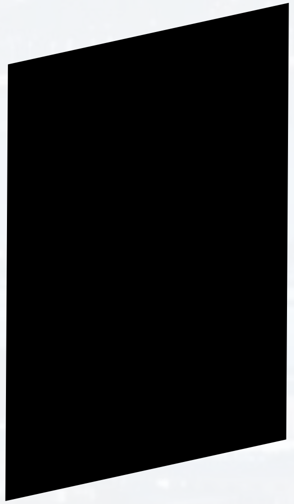
$$S = -\sum_{i}^{I} p_i \, ln p_i$$

- entropy high → information low
- entropy low → information high

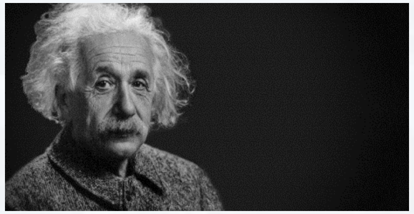
fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Come on, don't be so nerdy. It is an odd number.



This helps a lot: $p_2 = p_4 = p_6 = 0$ and thus, $p_1 = p_3 = p_5 = 1/3$



Machine Learning Algorithms:

Entropy is a mathematically precise measure of information!

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

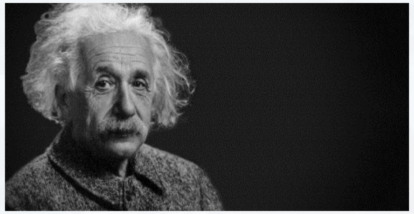
- entropy high → information low
- entropy low → information high

fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

Come on, don't be so nerdy. It is an odd number.

Hence, $\boldsymbol{S = -\frac{1}{3} ln\left(\frac{1}{3}\right) * 3 = 1.10 \ldots}$



Machine Learning Algorithms:

Entropy is a mathematically precise measure of information!

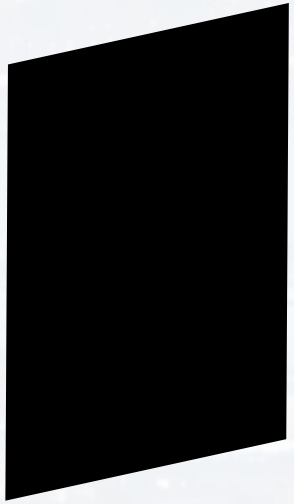
$$S = -\sum_{i}^{I} p_i \, ln p_i$$

- entropy high → information low
- entropy low → information high

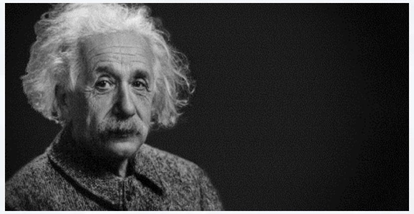
<u>fair die:</u> events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

I give up. It's a five. Next time I am gonna play with Schroedinger.



Fantastic! So all $p_n = 0$ except for n = 5



Machine Learning Algorithms: Entropy

Entropy is a mathematically precise measure of information!

$$S = -\sum_{i}^{I} p_i \, ln p_i$$

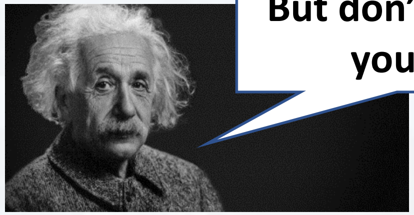
- entropy high → information low
- entropy low → information high

fair die: events are 1, 2, 3, 4, 5, 6

p_n are $p_1, p_2, p_3, p_4, p_5, p_6$ = 1/6 for all p_n

I give up. It's a five. Next time I am gonna play with Schroedinger.

Hence, $S = -0 \, ln(0) * 5 - 1 * ln(1) = 0$

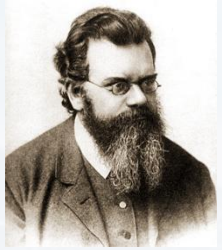


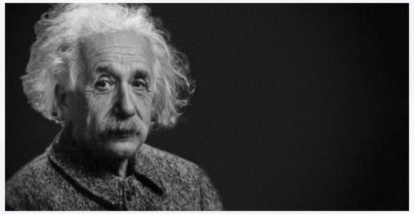
But don't mention your cat!

Entropy is a mathematically precise measure of information!

$$S = -\sum_{i}^{I} p_i \, lnp_i$$

- entropy high → information low
- entropy low → information high





Information:	Entropy:
none (any number)	S = 1.79
not a six	S = 1.61
an odd number	S = 1.10
the actual number	S = 0.00

M. Hohle:

Thank you for your attention!