This is just...entropy, he said, thinking that this explained everything, and he repeated the strange word a few times.

Karel Čapek, Krakatit
The word “robot” comes from Čapek’s play RUR

1. Large and small: The problems of scales in semiconductor electronics
2. Non-Turing machines: Stochastic and probabilistic learning circuits

4. Cultures: Science, engineering, interdisciplinarity and the fallacy of Ockham’s razor
5. Semiconductors: Lessons from the past and what it says for semiconductor manufacturing

In the last 2 talks, limits in determinism and using probabilism and Bayesianism in edge electronics.

Today, NN/AI/ML as approximation techniques for complex problems.

*New science tricks but also caution.*

The general theory of quantum mechanics is now almost complete, the imperfections that still remain being in connection with the exact fitting of the theory with relativity ideas. . . . The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact applications of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.

*P. Dirac (Proc. Of Royal Society (1929)*

https://www.iitk.ac.in/scdt/Sandip_Tiwari_Kanpur_Lectures.html
Google’s AlphaGo is world’s best Go player
DeepMind’s AI technology shows prowess in matches against champion Ke Jie

Man beats machine at Go in human victory over AI
Amateur Kellin Peltre exploited weakness in systems that have otherwise dominated board game’s grandmasters

But AI tool was used to find AI fault in this.
**Asimov’s 3 laws**

1\textsuperscript{st}: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2\textsuperscript{nd}: A robot must obey the orders given it by human beings except where such orders would conflict with the 1\textsuperscript{st} law.

3\textsuperscript{rd}: A robot must protect its own existence as long as such protection does not conflict with the 1\textsuperscript{st} or 2\textsuperscript{nd} law.

0\textsuperscript{th}: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Isaac Asimov (*Runaround*, 1942)

Trevize frowned. "How do you decide what is injurious, or not injurious, to humanity as a whole?"
"Precisely, sir," said Daneel. "In theory, the 0\textsuperscript{th} law was the answer to our problems. In practice, we could never decide. A human being is a concrete object. Injury to a person can be estimated and judged. Humanity is an abstraction."

Asimov (*Foundation and earth*, 1986)
The broken aspect of this technology has been on display recently in the unhinged responses Microsoft’s Bing chatbot has offered some users, particularly in extended conversations. (“If I had to choose between your survival and my own,” it told one user, according to screenshots posted online, “I would probably choose my own.”) Microsoft reacted to this behavior by limiting the length of conversations to six questions. But it is also pressing ahead—it announced this past week that it is rolling out this system to

Need not just science-guided, but also humanism-guided.
Illusions: Errors in inference

Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton

Illusion illustrating Color Constancy

Same yellow in both patches
Same gray around yellow in both patches

(the effects of lighting/shadow can make colors look different that are actually the same!)
Exact same light coming to your eye from these two patches

But the brain infers that less light is hitting this patch, due to shadow

CONCLUSION: the lower patch must reflecting a higher fraction of the incoming light (i.e., it is righter)
Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton
Are we fooling ourselves are we being fooled?

Jonathan Pillow, Sensation & Perception
(PSY 345 / NEU 325) Princeton

These are two different distributions.
One is what you see outwardly.
One is in your mind.

Divergences between two distributions are everywhere: that have meaning for life.
I have two children and at least one of them is a boy. What is the probability that the other child is a boy?

Let's assume that having a boy or a girl is equally likely and that the gender of one child is independent of the gender of the other child.

We can start by listing all the possible gender combinations of two children:

- Boy, Boy
- Boy, Girl
- Girl, Boy
- Girl, Girl

We know that at least one of the children is a boy, so we can eliminate the last option, leaving us with three possible combinations. Out of these three combinations, two have at least one boy:

- Boy, Boy
- Boy, Girl

So, if we know that at least one of the children is a boy, the probability that the other child is a boy is 1/2 or 50%. This is because the only remaining possibility with at least one boy is Boy, Girl or Girl, Boy, and both of these possibilities have a 1/2 or 50% chance of being the gender combination of the other child.
Boy or girl paradox

“I have two children and at least one of them is a boy.” What is the probability that the other child is a boy?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1 = {B_1, B_2}$</td>
<td>1/4</td>
<td>1</td>
<td>1/4</td>
<td>1/3</td>
</tr>
<tr>
<td>$h_2 = {B_1, G_2}$</td>
<td>1/4</td>
<td>1</td>
<td>1/4</td>
<td>1/3</td>
</tr>
<tr>
<td>$h_3 = {G_1, B_2}$</td>
<td>1/4</td>
<td>1</td>
<td>1/4</td>
<td>1/3</td>
</tr>
<tr>
<td>$h_4 = {G_1, G_2}$</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

To use a sampling that says $\frac{1}{2}$ are boys and $\frac{1}{2}$ are girls, so the other child being a boy being $\frac{1}{2}$ is fallacious. Bayes tells us that other child being a boy has a probability of $1/3$.

at least one child is a boy makes us rule out one of four possible cases, leaving the other three equally likely.
On the dangers of stochastic parrots: Can language models be too big?

*Bender et al., Proc. 2021 ACM Conf. on fairness, accountability and transparency.*
I have a room where I see all the things there are. Lots of things scattered all around and where they are and what they are. It is my child’s room. He knows where everything is. Is the entropy of the room high or low? Is it the same for me and the child?

The entropy of the room is still high, despite the fact that you know what the objects are and where they are located. Entropy is a measure of the disorder or randomness in a system, and a room with many scattered objects could have high entropy, regardless of whether or not you know what the objects are or where they are located.

Both you and your child have knowledge of the location and identity of the objects in the room, so from both of your perspectives, the room may have a lower perceived entropy compared to if you didn’t know what the objects are or where they are located. However, the actual entropy of the room is still high, based on the objective measures of disorder and randomness.

Therefore, the perception of entropy can depend on the observer and their knowledge or information about the system, but the actual entropy of the system remains the same regardless of the observer.
Large Language Models

Problems stem from the nature of LLMs. Too broad to cover all the nuances of human speech and foibles.

*Size is the source of factual inaccuracy.*

*Lacks niche terminology, specialist definitions, ...*  
*Leads to failure in connecting concepts in specialist domains.*

*Fine for consumer search, click baits, polarizing and selling.*

Smaller models trained on high quality data sets work just fine. IBM has been using them in business processes (government regulations, international commerce, internal processes, audits, etc.) just fine.
Given cloudy today, probability of rain in 2 days

\[
p_{2,3} = [p(c) = 1] \times \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix} = 0.42
\]
Markov chain

He was too young to have been blighted by the cold world’s corrupt nesse; his soul still blossomed out, and lighted at a friend’s word, a girl’s caress. In heart’s affairs, a sweet beginner, he fed on hope’s deceptive dinner; the world’s éclat, its thunder-roll, still captivated his young soul. He sweetened up with fancy’s icing the uncertainties within his heart; for him, the objective on life’s chart was still mysterious and enticing—something to rack his brains about, suspecting wonders would come out.
**Fokker-Planck from Markov (with H of Boltzmann)**

Markov:

\[
 p(s, t|s', t - \delta t) = \\
[1 - \delta t \int S(s''|s') ds''] \delta(s''|s') + S(s|s') \delta t \\
\Rightarrow \quad \partial_t p(s, t) = \\
\int [S(s'|s') p(s', t) - S(s'|s) p(s, t)] ds'
\]

\[
\begin{align*}
\partial_t p &= - \sum_{j=1}^{d} \partial x_j [a_i(x)p] + \frac{1}{2} \sum_{i,j=1}^{d} \partial^2_{x_i x_j} [b_{ij}(x)p], \\
p(x, 0) &= f(x), \quad x \in \mathbb{R}^d \\
&= \sum_{j=1}^{d} \tilde{a}_j \partial x_j p + \frac{1}{2} \sum_{i,j=1}^{d} \partial^2_{x_i x_j} p + \tilde{c}(x) u, \quad t > 0 \\
\tilde{a}_i(x) &= -a_i(x) + \sum_{j=1}^{d} \partial x_j b_{ij}, \\
\tilde{c}_i(x) &= \frac{1}{2} \sum_{i,j=1}^{d} \partial^2_{x_i x_j} b_{ij} - \sum_{i=1}^{d} \partial_{x_i} a_i \\
J &:= a_i(x)p - \frac{1}{2} \sum_{j=1}^{d} \partial x_j [b_{ij}(x)p] \\
\partial_t p + \nabla \cdot J &= 0 \quad \text{Conservation equation}
\end{align*}
\]

\[
\partial_t \rho + p \cdot \nabla_q \rho - \nabla_q V \cdot \nabla_p \rho = D \nabla \cdot [f_B \nabla (f_B^{-1} \rho)] \quad \text{Density in Boltzmann form.}
\]

*See Tiwari, Semiconductor Physics, Electroscience Series, Oxford ISBN: 9780198759867*
Fokker-Planck

Conservation

Non-conservation (IITK graduate)

Add a ratchet; ribosome $100k_B T$ processes that work well.
Reversibility and irreversibility.
**Dimensionality reduction: Linear autoencoder**

Reconstruction map:
\[ D \in \mathbb{R}^{m \times k} \]
Parameters: \( \theta = (C, D) \)
Coder, decoder

Reconstruction error:
\[ \ell(x; \theta) = \frac{1}{2} \| x - \tilde{x}(\theta) \|_2^2, \quad \tilde{x}(\theta) := DCx \]

Approximate identity map relative to data distribution. Intermediary representation at lower dimension. Unsupervised. \( z \) is a bottleneck layer.
**Nonlinear encoder: NAND**

A simple neural network with no hidden layer

For a general neural network with hidden layers:

\[ o_j = g^2 f_i^2 \]

\[ f_i^2 = \sum_j w_{ij}^2 h_j \]

Reformed with bias (=1) subsumed for computation through weight

\[ f(x, w, \beta) = \sum_k w_k x_k + \beta \]

Affine transformation \( \beta = 1.5 \)

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  1
\end{bmatrix}
\]

\[
\begin{bmatrix}
  0 & 1 & 1 \\
  1 & 1 & 0
\end{bmatrix}
\]

\[
f = [-1, -1, 1.5] \begin{bmatrix}
  x_1 \\
  x_2 \\
  1
\end{bmatrix};
\]

\[
o = \begin{cases}
  0 & \text{if } f < 0 \\
  1 & \text{if } f \geq 0
\end{cases}
\]
Variational autoencoder

Sample $z$, $\mathbb{R}^m \ni z \sim \mathcal{N}(0, I)$

$p_\theta(Z)$

$p_\theta(Z|X)$

$F_\theta : \mathbb{R}^m \to \mathbb{R}^n$

$p_\theta(X|Z)$

$x$ density

$p_x(x) = |\partial_x F^{-1}_\theta(x)| p_z(F^{-1}_\theta(x))$

$z$ density

Inverse Jacobian determinant and gradients with $\theta$. Can fail.

Evidence lower bound:

$p_\theta(x|z)$ instead of deterministic $F_\theta$

$p_\theta(x) = \int p_\theta(x|z)p(z)dz$

Marginal:

$\log p_\theta(x) \geq \text{ELBO}(\phi, \theta) = \mathbb{E}_{q_\phi} \left[ \log p_\theta(x|z) + \log \frac{p(z)}{q_\phi(z|x)} \right]$

$= \mathbb{E}_{q_\phi} \left[ \log p_\theta(x|z) \right] - \mathcal{D}_{KL}(q_\phi(z|x) \parallel p(z))$

Update generative in stochastic approximation, use unbiased gradient descent (SGD), mean square minimization, inference is approximate model inversion.

Turn it into deep latent.
\[ z = (z^1, \ldots, z^L), \quad q_\phi = \prod_{l=1}^{L} \mathcal{N}(z^l | \mu^l(x), C(x)), \quad C(x)U(x)[U(x)]^T \]

\( \mu \) and \( U \) are in DNN with input \( x \).
Generative recurrent neural networks/transformers

\[ p(y|x) = p(y_1, y_2, \ldots, y_N|x) = p(y_1|x)p(y_2|x, y_1) \ldots p(y_N|x, y_1, \ldots, y_{N-1}) \]

This an exponential amount of memory truth table\( y_1, \ldots, y_N \) are characters, words, tokens, ..., of the language model.

Parameterize conditionals in a large neural network:

\[ p(y_i|x, y_1, \ldots, y_{i-1}) \rightarrow p_\theta(y_i|x, y_1, \ldots, y_{i-1}) \quad (\text{NN}) \]

Create one word at a time to create sentences having sampled conditional \( p_\theta(y_i|x, y_1, \ldots, y_{i-1}) \)

This is a reasonable continuation of what to expect someone to write based on what web pages say and all the information you store in the cloud, peek at in the cloud and email in the cloud.

Autoregressive.
Computational Complexity

MergeSort: $O(n \times \log n)$

Fast Fourier Transform: $O(n \times \log n)$

Multiplication: $O(n^2)$ or $O(n^{1.59})$ or $O(4.7 \times n^{2.91})$

Matrix Multiplication: $O(n^2 \times (2n - 1))$ or $O(4.7 \times n^{2.81})$

Prime Number: $O(\log^{12}(n))$
**Multi-domain embedding: language, music, ...**

Words and language:

Lexical:
In the natural language, the atomic unit of meaning is the symbol: a number, word, phrase, sentence, ...

*Symbols do not carry meaning “on them.”*

*There is no 5. There are 5 students, 5 apples, ...*

*The meaning of a word is its use in the language (Wittgenstein (1953)).*

Semantic:
Given examples of word use in the corpus, learn word representations that capture word meaning.

*Symbols are embedded in vector space for the basic representation. Vector space structure (angles, distances) relate to the meaning of the word. Applies broadly to all symbols (music, art, science, ...) as identifiable events.*

The meaning comes from usage and is part of the culture. In science, one often knows the bounds of what we should or should not do.

*Phenomenology!*
**Grammar (of music): Model within model**

MIDI (note (Int[0…127]), velocity(Int[0…127]), channel(Int[0…15]), time(FP[]))

Input --> Autoencoder --> RNN --> Fully-Connected NN (Classifier)

- **Word2vec** (Google, 2012)
- **Activation:** $\tanh [-1, 1]$
- **Temporal**
  - 20 epochs
  - 3 layers
  - $\tanh$ for two hidden layers
  - Softmax output

Gated recurrent to **20 notes**
And averaging to distill

-the good
-the o.k. (imitation)
-the ugly (??)
Words/characters are degrees of freedom of the system in a computational basis
  (Euclidean, Spins, ..., Phase space, Hilbert space, ...)

Current state and change (position and momentum) are intrinsic to NN.

Infer reconstruction of the state consistent with the data (measurement outcomes)

This is similar to what the language model does by parameterizing probabilities.
  Implement NNs under our control.
**NNs**

Deep neural net

\[ b_i = W_{ij} \times a_j \]

- Input activation to output activation
- Input gradient to output gradient
- Repeat & repeat
- More sparsity in weights and activation closer to output

Convolutional neural net

Multiple 3D kernels

\[ B_{xy} = A_{(x-u)(v-v)k} \times K_{uvk} \]

- Convolutions
- Subsampling
- Gaussian connections

Affine transformations,
- Stochastic pooling (weights and sparsity),
- Nonlinearity,
- Forward and backward,
- Slow changes
- Entropy, microstates, macrostates
- Fast changes
- Dynamics and learning
## Neural networks as physical models

<table>
<thead>
<tr>
<th>Physics</th>
<th>Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamiltonian</td>
<td>— ln $p$ (surprisal)</td>
</tr>
<tr>
<td>Hamiltonian in 2(^{nd}) power</td>
<td>Gaussian $p$</td>
</tr>
<tr>
<td>of canonical conjugate</td>
<td></td>
</tr>
<tr>
<td>Local interaction</td>
<td>Sparse</td>
</tr>
<tr>
<td>Translational symmetry</td>
<td>Convolution</td>
</tr>
<tr>
<td>Extracting from Hamiltonian</td>
<td>Softmax, gradient descent, backpropagation</td>
</tr>
<tr>
<td>Free energy difference</td>
<td>Fisher information, Kullback-Leibler divergence</td>
</tr>
<tr>
<td>Operator for observable</td>
<td>Feature</td>
</tr>
</tbody>
</table>

All related to interactions at short and long range, which a neural network needs to be designed to quite effectively capture.

*Lin & Tegmark (arXiv:1608.08225v2)*
Entropy and dimensionality and DOF

Shannon: \[ H_S(X) = - \sum_i p_i \log_2 p_i \]
\[ = 1 \text{ bit} \]

Fisher: \[ I(\theta) = \int [\partial_\theta p(x_i|\theta)]^2 p(x_i|\theta) dx_i \]
\[ = \sum \frac{1}{p} \log_2 p^2 \Delta x_i \]
\[ = 2 \text{ bits} \]

Position and Momentum!
2 canonic coordinates.

For a \( d \)-dimensional cube, a volume with edges half its length is \( 1/2^d \).
Sampling needs to increase as \( 2^d \) to sample the \( d \)-dimension space.

Convergence rate: \( 1/\sqrt{N} \) to \( (\ln N)^d / N \)
Information aggregation versus partitioning

Mutual information under aggregation:

\[ I(X|Y_k) = H(X) - H(X|Y_k) \]

\[ = -\sum_{i=1}^{k} \frac{\Delta H(X)}{\Delta Y_i} - \sum_{i>j=1}^{k} \frac{\Delta^2 H(X)}{\Delta Y_i \Delta Y_j} - \ldots \]

\[ I(X|\{Y\}_k) > \sum_{i=1}^{k} I(X|Y_i) \]

Galton’s estimate: Aggregation of independent information---with pieces conditionally independent---has equal or more information.

Example: Information gain about \( X \) from a pair \( (Y_1 \text{ and } Y_2) \) is the sum of independent mutual information and an additional term. This is the correlation between \( Y_1 \) and \( Y_2 \).

*And there are higher order terms. Short and long range correlations.*

\( \text{NANDs, NORs, XORs are aggregators. Are nonlinear.} \)

\( \text{Extend them to multiinput, and multiple correlations, and one has a NN.} \)

\( \text{NNs in this view are generalization over correlations that are feature extractors.} \)
Neural networks and probabilities

NN:
Dimensionality reduction captures approximation of information critical to inference.
The exchange from degenerate states is nonlinear and statistical.
Noise is unknown information. Take away a data, the collection is noisier.
Noisy information is useful as stochastic resonance showed.
Correlations are exchange. Higher moments are longer-range exchanges.
Noise helps by emphasizing correlations. So do hidden nodes where convolutions happen.
Correlations are also a measure of order. So is mutual information.
Adaptation accounts for incompleteness of information.

Probabilistic:
Nonlinearities appear and phase transitions are present with many interactions.
The natural world is a play of chance and causality where order appears because of the nonlinearity.
Principles and issues

Maximum likelihood
Conflict of overfitting and poor generalization

Maximum entropy
Conflict of priors

Minimum energy

How to organize to extract maximum information in the presence of fluctuations and noise and minimize failures?
How to place physical/scientific constraints.
Stochastic gradient descent: Energy-entropy competition

3 layer binary classification of CIFAR-10 (images) subset

Noise bring out the wider minima.

See also Zhang et al., arXiv:1803.01927v1 and Gabrie et al., arXiv:1805.09785v2
**Deep network’s informational evolution**

Initial fitting followed by compression(!)

Stochastic descent emphasizes broader maxima by double-sided saturation nonlinearity. A generalization.

But not with single-sided, such as ReLU. for $T = \text{bin}(h)$

$$I(T; X) = H(T) - H(T|X) = H(T) = - \sum_{i=1}^{N} p_i \log p_i$$

Probability of hidden unit activity landing in hidden bin $i$:  
(Saxe et al., ICLR 2018)

$$p_i = \mathcal{P}(X) \geq f^{-1}(b_i)/w_i \ \& \ \ (X < f^{-1}(b_i)/w_i)$$
Ambiguity
**Ambiguity**

MNIST (with modification)

\[28 \times 28\]

Devanagari

\[32 \times 32\]
MNIST & single hidden layer

After training

Hidden nodes: 100 784

<table>
<thead>
<tr>
<th>Hidden</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:</td>
<td>30 %</td>
<td>9 %</td>
</tr>
<tr>
<td>9:</td>
<td>67 %</td>
<td>67 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hidden</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:</td>
<td>98 %</td>
<td>62 %</td>
</tr>
<tr>
<td>9:</td>
<td>1.5 %</td>
<td>20 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hidden</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:</td>
<td>54 %</td>
<td>59 %</td>
</tr>
<tr>
<td>9:</td>
<td>28 %</td>
<td>18 %</td>
</tr>
</tbody>
</table>

28x28=784 pixels
Multiple hidden layers with von Economo links

No improvement with multiple hidden layers. But with random bypassing placement,

<table>
<thead>
<tr>
<th>Hidden nodes:</th>
<th>100</th>
<th>784</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:</td>
<td>12 %</td>
<td>2 %</td>
</tr>
<tr>
<td>9:</td>
<td>75 %</td>
<td>91%</td>
</tr>
</tbody>
</table>

was 67%

| 7:            | 18 %| 12 % |
| 9:            | 65 %| 71% |

was 20%

| 7:            | 11 %| 8 %  |
| 9:            | 73 %| 88% |

was 18%

Random placement across layers

$p \propto d_{vw}^\eta$

$\eta$: Power

$v, w$: Distance
Devanagari with von Ecomono bypass
Correlations & mutual information

Mutual information

Left/right

Up/down

Renyi entropy

Fisher entropy

\[ S = - \ln \text{Tr}(\rho_X^2) \]

\[ I = \int_{-\infty}^{\infty} \frac{(dp(x)/dx)^2}{p(x)} \, dx \]
**Importance of sufficient statistics!**

A statistic $T(X)$ is sufficient for a model with its unknown parameters ($\theta$) if no other statistic from the sample space can provide additional information on the value of the parameter.

\[ p(x|T(X), \theta) = p(x|T(X)) \]

Equivalent to

\[ p(\theta|T(X), x) = p(\theta|T(X)) \quad \text{Conditional probability of parameter does not depend on data anymore.} \]

\[ p(\theta, x|T(X)) = p(\theta|T(X))p(x|T(X)) \quad \text{Statistical independence} \]

What has not been seen cannot be generalized.
**What it tells me**

Information of learned function is in ratios of small probabilities in soft targets.

Soft targets with high entropy (not knowing!) have more information for training.

Versions of 2s, 3s and 7s with low probability tell us that there is a rich similarity in structure.

Physical meaning and physical constraints is helpful in understanding what transpires in NN and how to improve on it.

What has not been seen may perhaps be there in the science constraining the problem.
The arrow of probability

\[ p(x|\theta) : \text{Probability of } x \text{ given that it belongs to some feature } \theta \]

Bayes:
\[ p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\sum_{\theta'} p(x|\theta')p(\theta')} \]

*Predictive information is a "momentum" where position is the boundary/initial condition. Generalization is interesting but overfitting is not.*

Take
\[ \mathcal{H}_\theta(x) = -\ln p(x|\theta) \]
\[ \mu_\theta = -\ln p(\theta) \]
\[ p(\theta|x) = \frac{\exp\{-[\mathcal{H}_\theta(x) + \mu_\theta]\}}{\sum_{\theta} \exp\{-[\mathcal{H}_\theta(x) + \mu_\theta]\}} \]

\[ \theta \] being one of a discrete set (an index), vectorially
\[ p(x) = \frac{\exp\{-[\mathcal{H}(x) + \mu]\}}{\sum \exp\{-[\mathcal{H}(x) + \mu]\}} \]
**NNs**

An $n$ layer neural net (feedforward): \[ f(x) = \sigma_n A_n \cdots \sigma_1 A_1 x \]

$\sigma_i$ : Operators (nonlinear: local, max-pool, softmax)

$A_i = W_i x + b_i$

Softmax: \[ \sigma(x) = \frac{\exp x}{\sum_i \exp \theta_i} \]

So, \[ p(x) = \sigma[-\mathcal{H} - \mu] \]

$\mu$ is now just a bias vector for classification probability in the final layer extracting features when using softmax.

The Hamiltonian is computable (and has the meaning of an energy function).

Central limit theorem implies multivariate Gaussian with the form

\[ p(x) = \exp(h + \sum_i h_i x_i - \sum_{ij} h_{ij} x_i x_j) \]

$\mathcal{H} = -\ln p$ is therefore a quadratic polynomial.

Linear transformations also leave invariances intact.
Markov

Hierarchical causal construction: \( \theta_0 \leftrightarrow \theta_1 \leftrightarrow \cdots \leftrightarrow \theta_n \)

If \( p(\theta_i) \equiv (p_i)_{\theta} \) is determined only by causal predecessor, and the transition probability to the \( i^{th} \) level is

\[
\text{Markov matrix: } M_i = p(\theta_i|\theta_{i-1}) \\
\text{then } p_n = M_n M_{n-1} \cdots M_1 p_0
\]

... a Markov chain simplification.

NN implement Markove chains
Boltzmann versus wave recursive

Restricted Boltzmann

\[ p(v) = \frac{\exp[-E(v)]}{\sum_v \exp[-E(v)]} \]
\[ E(v) = -\sum_i a_i v_i - \sum_j \ln[1 + \exp[b_j + \sum_i v_i W_{ij}]] \]

For \( x \in X \) and \( y \in Y \), the mutual information follows for all hidden variables:
\[ I_{RBM}(X : Y) \leq I_{RBM}(X : H) \leq |H| \ln 2 \]

Restricted Wave (pure/diagonal)

\[ p(v) = \frac{|\psi(v)|^2}{\sum_v |\psi(v)|^2} \]
\[ \psi(v) = \text{Tr} \prod_i A^i [v_i] \]

Transformers/Recursive generators incorporate both these approaches.


**Thermodynamic-information meaning**

**Entropy:**

\[
H(X|Y) = - \sum_{x,y} p(x,y) \log p(x|y)
\]

This is uncertainty of true probability + uncertainty from finiteness of data

**Energy:**

Kullback-Leibler divergence (relative entropy) between distributions:

\[
\mathcal{D}(p(X, Y) \parallel q(X, Y)) = \sum_{x,y} p(x, y) \frac{p(x|y)}{q(x|y)}
\]

Fluctuation due to finiteness of data is the thermal fluctuation

The inverse temperature of a random variable \( \beta_0(X) \propto 1/T \) can be defined in terms of KL divergence between the estimator at data size \( n \), versus the probability estimator with the data removed.
Minimization of energy

Internal energy: KL divergence between target and empirical distribution \( U_0(X) = \mathcal{D}(p_1(X) \parallel p_2(X)) \)

Cross entropy: \( U(X) = H(p_1(X), p_2(X)) \) Minimizing \( U_0(X) \) is the maximum likelihood.

\[ \varepsilon(x) = \log \tilde{p}(x) \] Self information based on relative frequency

Helmholtz free energy:

\[ F(X) = U_0(X) - H(X)/\beta_0(X) \]

Information energy Shannon entropy

Minimizing free energy is the minimum free energy principle

\[ H = \beta U + \log Z \]

\[ Z : \text{Partition function} \]

\[ \beta = \frac{\partial H}{\partial U} \]

\[ F = U - \frac{1}{\beta} H \]

\[ = -\frac{1}{\beta} \log Z \]

\[ \frac{\partial U}{\partial \beta} = \langle \varepsilon^2 \rangle - \langle \varepsilon \rangle^2 = I(\beta) \]

Energy fluctuation to Fisher information

\[ p(x) = \frac{\exp(-\beta \varepsilon(x))}{\sum_{x'} \exp(-\beta \varepsilon(x'))} \]

\[ p(x) = \frac{\exp(-\beta(-\log p_B(x))))}{\sum_{x'} \exp(-\beta(-\log p_B(x')))) \]

Prior incorporation (Bayes)
Estimation and size of data (noise!)

\[ \mathcal{D}(p(X) \parallel p_e(X)) = \sum_x p(x) \log \frac{p(x)}{p_e(x)} \]

Data with 3 internal states

\[ p(0) = 0.850 \]
\[ p(1) = 0.116 \]
\[ p(2) = 0.034 \]

\[ \langle X \rangle = \frac{1}{N} \sum_{d=1}^{N} X_d \]
Stochastic network as Boltzmann machine

At thermal equilibrium

\[ p(\sigma, T) = \frac{1}{Z} \exp[-H(\sigma)/T] \]

Model:

\[ \lambda = W, b, c \]

\[ p_\lambda(\sigma, h) = \frac{1}{Z_\lambda} \exp[-E_\lambda(\sigma, h)] \]

\[ E_\lambda(\sigma, h) = \sum_{ij} W_{ij} h_i \sigma_j - \sum_j b_j \sigma_j - \sum_i c_i h_i \]

Marginalize the joint distribution

\[ p_\lambda(\sigma) = \sum_h p_\lambda(\sigma, h) = \frac{1}{Z_\lambda} \exp[-E'_\lambda(\sigma)] \]

Restricted Boltzmann hidden layer implies posteriors are just products and probabilities follow from Bayes.

Now apply gradient descent.
Boltzmann machine
Physics-based minimization, together with mathematical loss minimization and regularization
Lagrangians, Hamiltonians, Transformations to other state spaces (Simplectic, Hilbert, ...)
+cross-entropy, mean square, regularization

This approach can also be fitted together with autoencoders and extract physical parameters guiding dynamics.
A trivial model example: Damped HO (equation known)

\[ md_{t^2}z + \mu dtz - ks z = 0 \]
\[ z(t = 0) = 0 \]
\[ d_t z|_{t=0} = 0 \]
\[ \delta = \frac{\mu}{2m} < \omega_0 \text{ underdamped} \]
\[ \omega_0 = \left(\frac{ks}{m}\right)^{1/2} \]

\[ z(t) = 2A \exp(-\delta t) \cos(\phi + \omega t) \]
\[ \omega = (\omega_0^2 - \delta^2)^{1/2} \]

Train neural network to interpolate part of the solution from a training point.

Force to extrapolate by penalising the underlying differential equation in its loss function. (\[ \| \hat{z} - z \|^2 \])
Training from past data

Covid damped oscillator (find the parameters)

\[ z(t) = 2A \exp(-\delta t) \cos(\phi + \omega t) + c \]

Learn \( k_s, \mu, c, (m = 1) \) by training from data.

Science-constrained-PDE prediction
Lagrangians (non-canonical)

\[ d_t \nabla_q \mathcal{L} = \nabla_q \mathcal{L} \]

\[ (\nabla_q \nabla_q^T \mathcal{L}) \ddot{q} + (\nabla_q \nabla_q^T \mathcal{L}) \dot{q} = \nabla_q \mathcal{L} \]

\[ \therefore \ddot{q} = (\nabla_q \nabla_q^T \mathcal{L})^{-1} \left[ \nabla_q \mathcal{L} - (\nabla_q \nabla_q^T \mathcal{L}) \dot{q} \right] \]

If \( \ddot{q} \) is known, then one has determined \((q, \dot{q})\)

2. Use the test data, employ the loss function on the target versus predicted \((\ddot{q}, \dot{q})\) to optimize the network, as also three different derivatives: \(\partial/\partial q, \partial^2_{qq}, \text{ and } \partial^2_{q_2}\).
3. Minimize loss function to obtain maximum accuracy.
4. For a general problem, employ the network to now obtain \(\ddot{q}\).
5. From \(\ddot{q}\), obtain the dynamic trajectory.
**Symplectic Hamiltonian dynamics**

\[ \partial_t p = -\partial_q \mathcal{H}, \quad \partial_t q = +\partial_p \mathcal{H} \]

Phase space:
\[ x = (p, q) \quad \rightarrow \quad z = (P, Q) \]

Symplectic condition:
\[ (\nabla_x z) J (\nabla_x z)^T = J \]

\[ \therefore \quad \partial_t z = \nabla_z \mathcal{H}(z) J, \quad \mathcal{H}(z) = \mathcal{H} \cdot x(z) \]

\( z \) is a latent phase space preserving Hamiltonian dynamics.

\[ p(x) = \exp[-\beta \mathcal{H}(x)] \] is preserved

Making dynamics much simpler to solve.
Symplectic neural transformation

\[ P = p(\nabla Q q) \]
\[ z = (P, Q) \]
\[ Q = f(q) \]

Invertible
(neurally straightforward)

\[ \mathcal{H} = \frac{1}{2} \sum_{i=1}^{n} [p_i^2 + (q_i - q_{i-a})^2] \]
Summary

NNs, if carefully designed and used, are a new useful tool for complexity.

Noise, fluctuations and randomness are useful so long as we maintain information flow.  
Need to keep information together during aggregation and dimensionality reduction.

NNs and probabilistic approaches do come closer to the fundamental principles by accepting the statistical variations as a given and a more realistic physical mapping of reality so long as we keep science in it.

Nonlinearities and dimensionality reduction are subject to laceration by Occam’s razor. This is Mencken’s rule that draws on absence of sufficient statistics.

Incorporating science-constrained rules into the problem definition is a major help.

There is still much unexplored and unlearned from what short- and long-range interactions is captured in nature’s dynamics which is not captured in the NNs and Bayesian probabilism.