Information Theory

Alexander G. Ororbia II

Introduction to Machine Learning

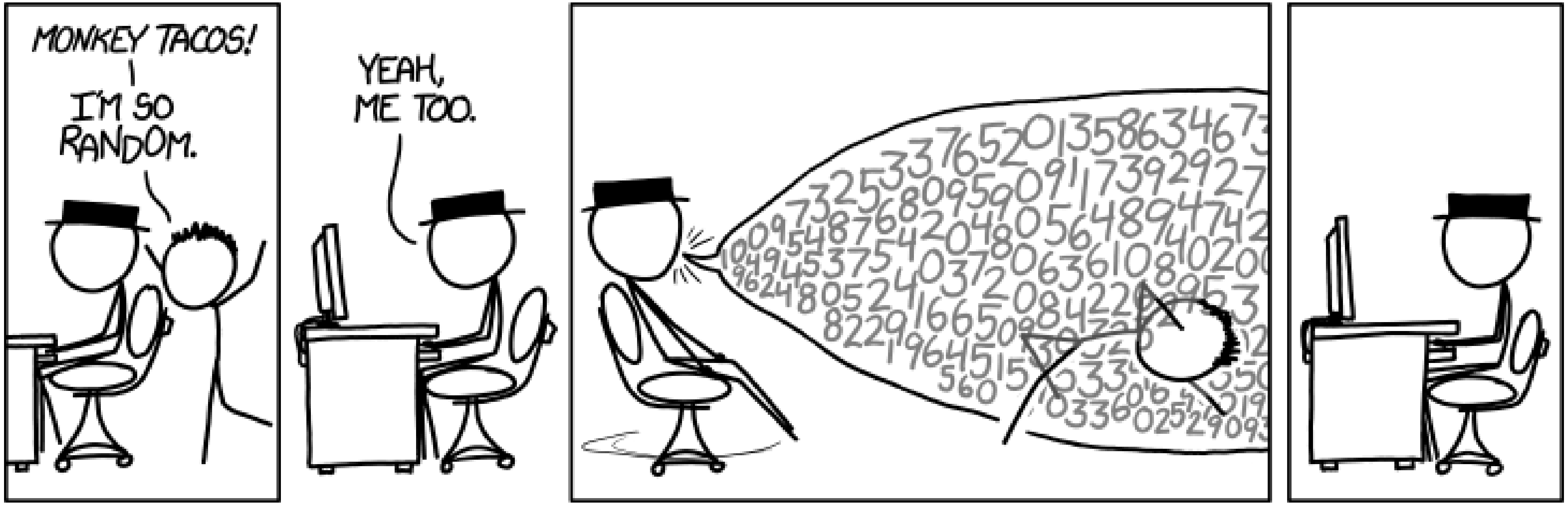
CSCI-635

9/8/2023

What is Information?

Information = minimum number of bits needed to encode all possible meanings of a given message, assuming all messages equally likely

What would be the minimum message to encode the days of the week field in a database? This is a type of ***compression***!



Information Theory (in the Classical Sense)

Model of innate information content of a thing, e.g., documents, images, messages, DNA

Information

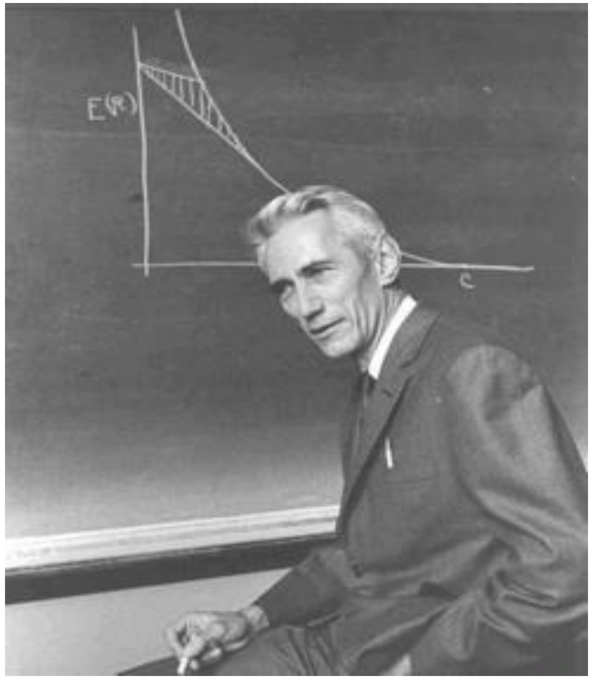
- That which reduces ***uncertainty***

Entropy

- A measure of information content
- **Conditional entropy**
  - Information content based on a context or other information

Information theory provides us with:

- Formal limitations on what can be: compressed, communicated, represented
- Information content can depend on probability of events & not just on number of outcomes
- Uncertainty = lack of knowledge about an outcome



Claude Shannon – founded information theory in 1948, also founded both digital computer & digital circuit theory in 1937

- 21-year-old master's student at MIT, wrote a thesis demonstrating that electrical application of Boolean algebra could construct / resolve any logical, numerical relationship = ***most important master's thesis of all time***

- Coined the term "***bit***"

Entropy: The Big Idea

- Entropy = average amount of information produced by a probabilistic stochastic source of data
- Example = a fair coin toss
  - If 10 tails in a row, we would be quite surprised (high entropy)
    - High entropy = something to be learned (this coin might be biased?)
  - Low entropy = what you expected, happened!
    (equal number of heads & tails)

A Note on Units:
Natural logarithms → *nats*
$\log_2()$ → *bits*

Information Measure

- How much information is received when we observe a specific value for a discrete random variable x?
- Amount of information is degree of surprise
  - Certain means no information
  - More information when event is unlikely
- Depends on probability distribution $p(x)$, a quantity $h(x)$
- If there are two unrelated events x and y we want $h(x,y) = h(x) + h(y)$
- Thus we choose $h(x) = - \log_2 p(x)$
  - Negative assures that information measure is positive
- Average amount of information transmitted is the expectation wrt $p(x)$ refered to as entropy

$$H(x) = -\sum_x p(x) \log_2 p(x)$$

Usefulness of Entropy

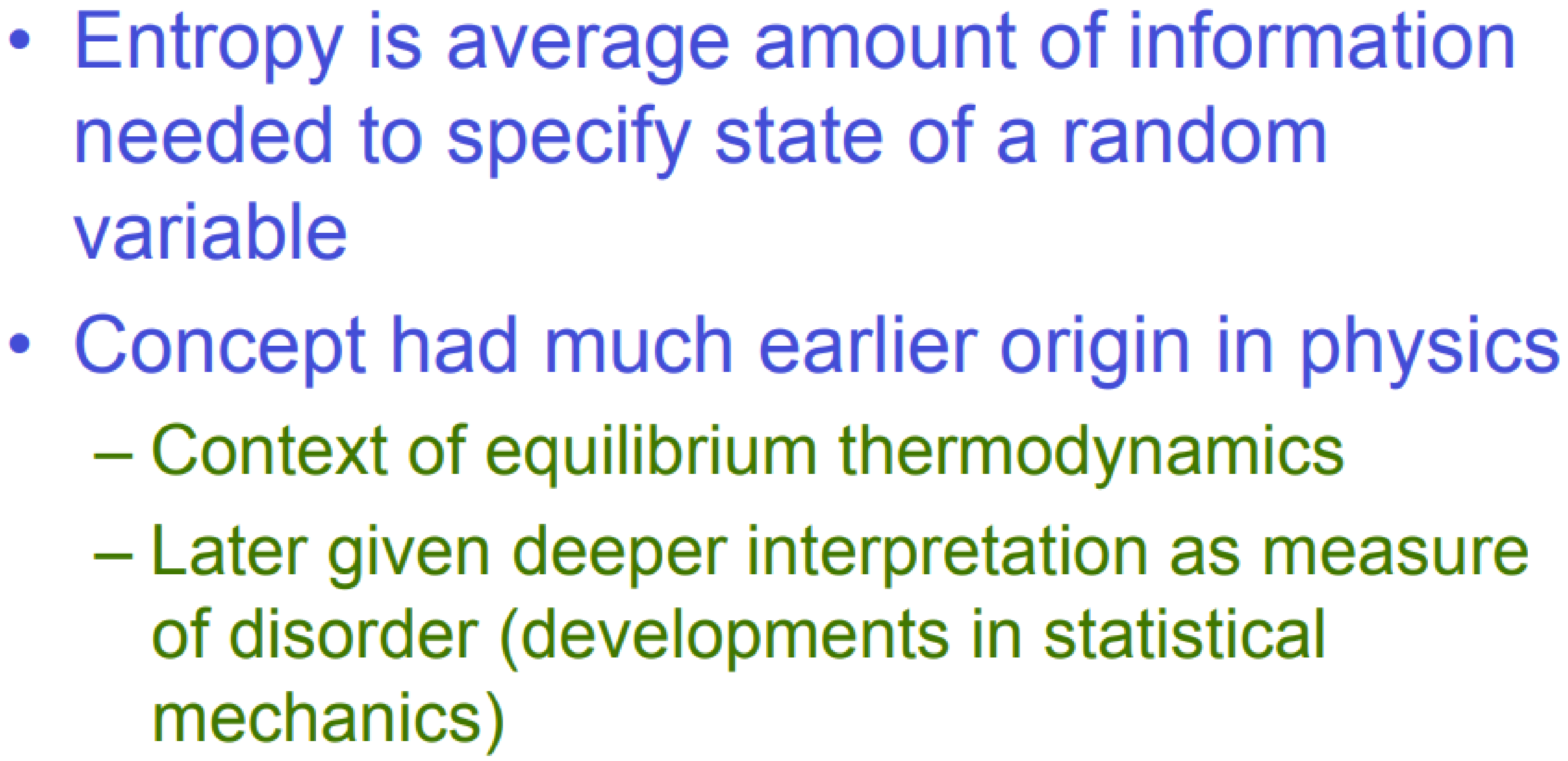
- Uniform Distribution
  - Random variable x has 8 possible states, each equally likely
    - We would need 3 bits to transmit
    - Also, $H(x)$ = -8 x $(1/8)log_2(1/8)=3\ bits$
- Non-uniform Distribution
  - If x has 8 states with probabilities

    $(1/2,1/4,1/8,1/16,1/64,1/64,1/64,1/64)$

    $H(x)=2\ bits$

- Non-uniform distribution has smaller entropy than uniform
- Has an interpretation of disorder

History of Entropy: Thermodynamics to Information Theory



- Entropy is average amount of information needed to specify state of a random variable
- Concept had much earlier origin in physics
  - Context of equilibrium thermodynamics
  - Later given deeper interpretation as measure of disorder (developments in statistical mechanics)

Masters of Entropy

World of Atoms

World of Bits

- Ludwig Eduard Boltzmann (1844-1906)
  - Created Statistical Mechanics
    - First law: conservation of energy
      - Energy not destroyed but converted from one form to other
    - Second law: principle of decay in nature– entropy increases
      - Explains why not all energy is available to do useful work
  - Relate macro state to statistical behavior of microstate
- Claude Shannon (1916-2001)
- Stephen Hawking (Gravitational Entropy)

Relative Entropy

- If we have modeled unknown distribution $p(\mathrm{x})$ by approximating distribution $q(\mathrm{x})$
  - i.e., $q(\mathrm{x})$ is used to construct a coding scheme of transmitting values of x to a receiver
  - Average additional amount of information required to specify value of x as a result of using $\mathrm{q(x)}$ instead of true distribution $\mathrm{p(x)}$ is given by relative entropy or K-L divergence
- Important concept in Bayesian analysis
  - Entropy comes from Information Theory
  - *K-L Divergence*, or *relative entropy*, comes from Pattern Recognition, since it is a distance (dissimilarity) measure

Relative Entropy or K-L Divergence

- Additional information required as a result of using $q(\mathrm{x})$ in place of $p(\mathrm{x})$

$$KL(p\,||\,q) = -\int p(x)\ln q(x)dx - \left(\int p(x)\ln p(x)dx\right)$$

$$= -\int p(x)\ln\left\{\frac{p(x)}{q(x)}\right\}dx$$

- Not a symmetrical quantity: $KL(p||q) \neq KL(q||p)$
- K-L divergence satisfies $KL(p||q)>0$ with equality iff $p(x)=q(x)$

Kullback Leibler (KL) Divergence: Some Nice Formulae

Gaussian KLD:

$$KL(p, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

$$p \sim N(\mu_1, \sigma_1)$$
$$\mathrm{q} \sim N(\mu_2, \sigma_2)$$

Bernoulli KLD:

$$KL(p||q)_{Ber} = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}$$

Mutual Information

- Given joint distribution of two sets of variables $p(\mathrm{x},\mathrm{y})$
  - If independent, will factorize as $p(\mathrm{x},\mathrm{y})=p(\mathrm{x})p(\mathrm{y})$
  - If not independent, whether close to independent is given by
    - KL divergence between joint and product of marginals

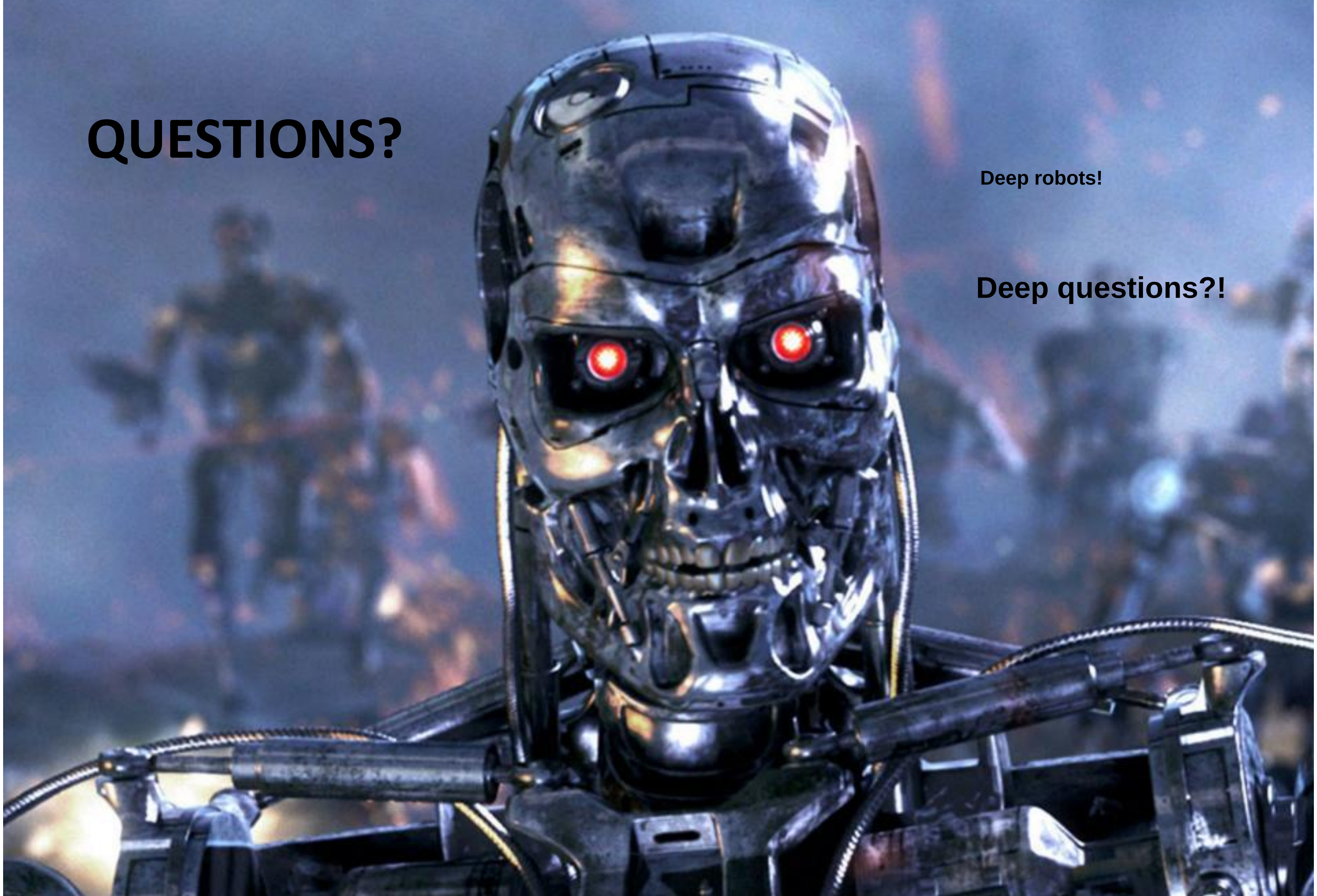
$$I[\mathrm{x},\mathrm{y}] = KL(p(\mathrm{x},\mathrm{y}) \,||\, p(\mathrm{x})p(\mathrm{y}))$$

$$= \iint p(\mathrm{x},\mathrm{y}) \ln\left(\frac{p(\mathrm{x})p(\mathrm{y})}{p(\mathrm{x},\mathrm{y})}\right) d\mathrm{x}\,d\mathrm{y}$$

- **Measures the correlation of two random variables!**

- Called Mutual Information between variables x and y

QUESTIONS?



Deep robots!

Deep questions?!

Claude Shannon 1948

Shannon defined information in a way not previously used

Shannon noted that information content can depend on the probability of the events, not just on the number of outcomes.

Uncertainty is the lack of knowledge about an outcome.

Entropy is a measure of that uncertainty (or randomness)

- in information
- in a system

Fundamental Questions Addressed by Information Theory

What is the ultimate data compression for an information source?

How much data can be sent reliably over a noisy communications channel?

How accurately can we represent an object (e.g. image, etc.) as a function of the number of bits used.

Good feature selection for data mining and machine learning