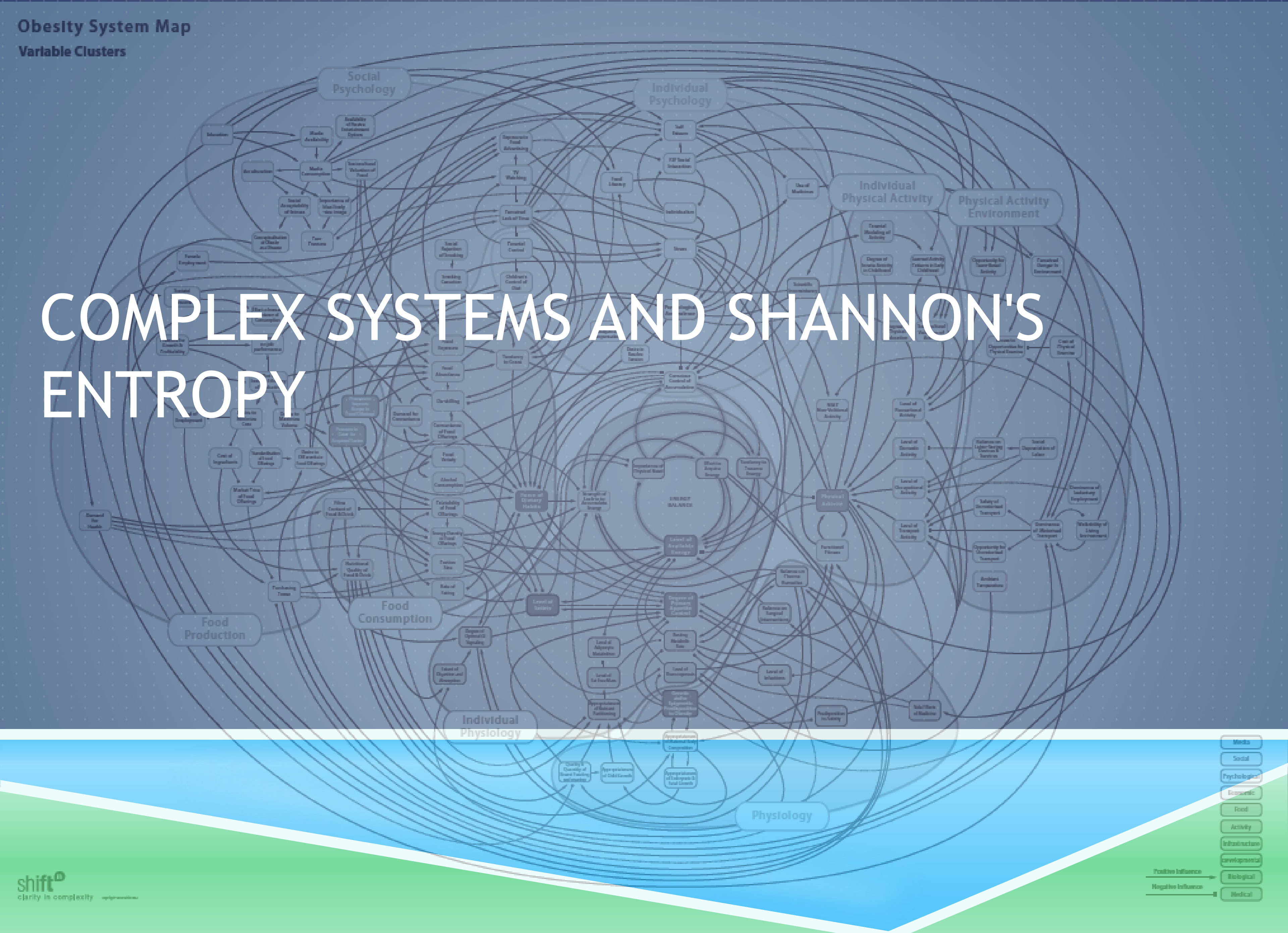
COMPLEX SYSTEMS AND SHANNON'S ENTROPY



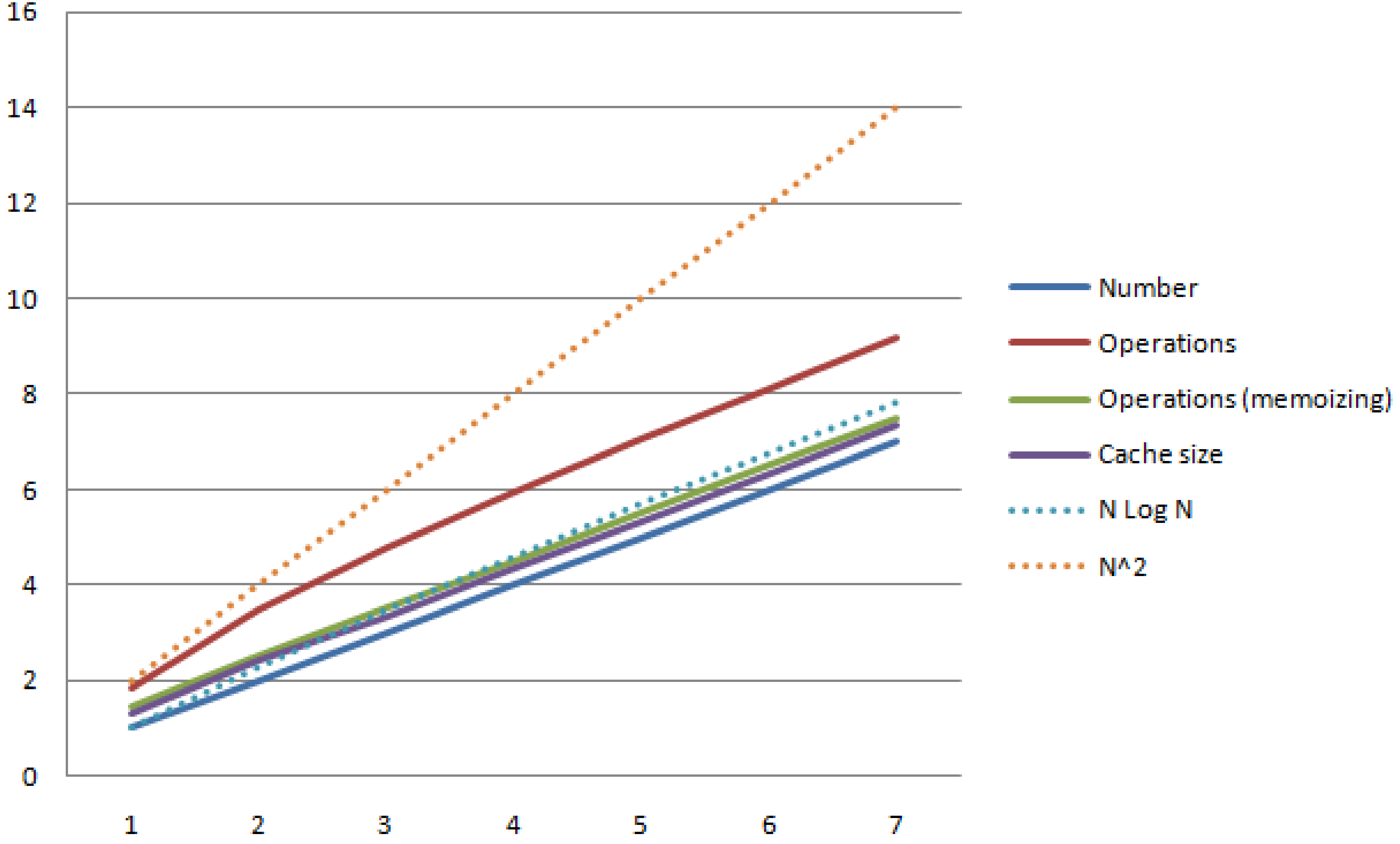
COMPLEX SYSTEMS

- What are complex systems?
  - Systems comprised of a (usually large) number of (usually strongly) interacting entities, processes or agents*, the understanding of which require the use of or development of new scientific tools, non- linear models, out of equilibrium descriptions and computer stimulations.
    - Advances in Complex Systems Journal
  - That is, how parts of a system give rise to the collective behaviors of the system, and how the system interacts with its environment.

*Agents can be anything between a subroutine and a conscious entity, generally for something to be called an agent, it must have some form of autonomy .

"Science is organized knowledge. Wisdom is organized life."
- Immanuel Kant

Log scale of time and space complexity



TYPES OF COMPLEX SYSTEMS

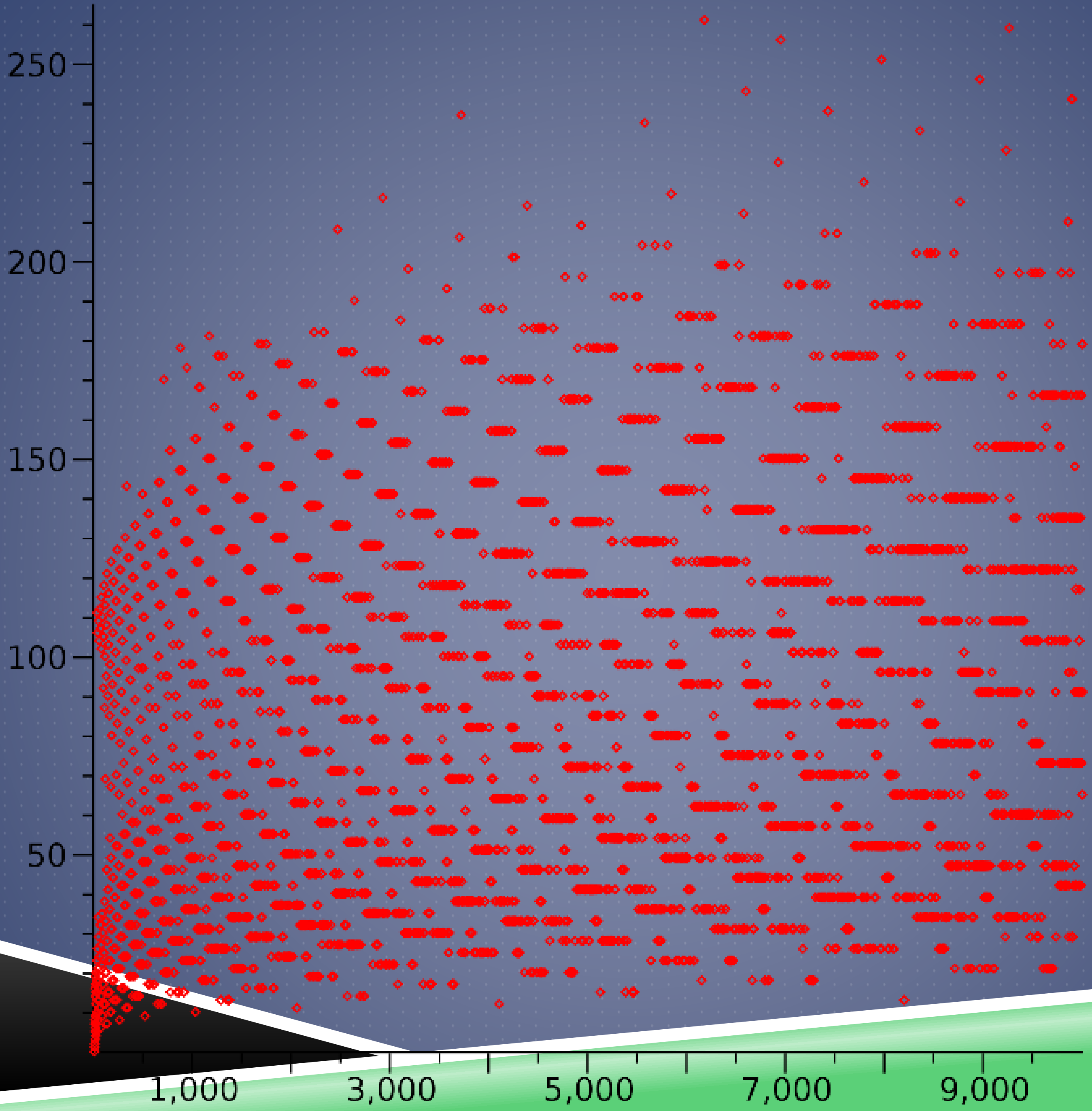
- There are two main types of complex systems:
  - Systems that are made complex by a large number of components with complex relationships and/or interactions.

    For example: World Economy, Fluid flow/ avalanche, Bandwidth on the internet

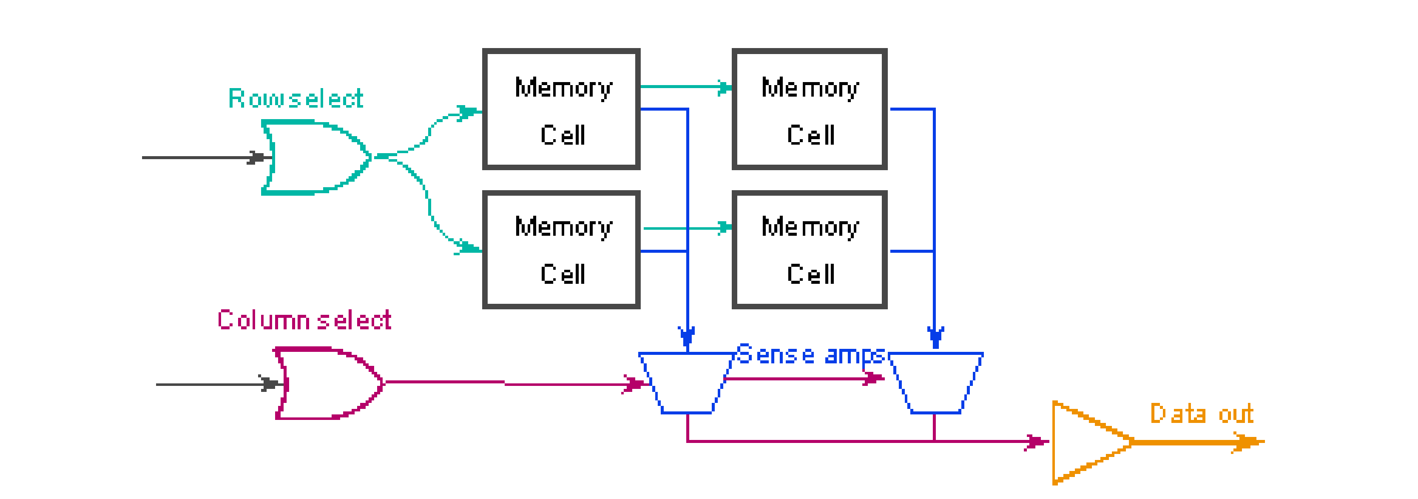
  - Systems that have an inherit non-linear behavior, even if they can be easily or concisely described.

    For example: some non-linear oscillators, non-linear and chaotic systems, Collatz iteration (3n-1) and of course, weather systems.

"Not everything that can be counted counts, and not everything that counts can be counted."
Albert Einstein

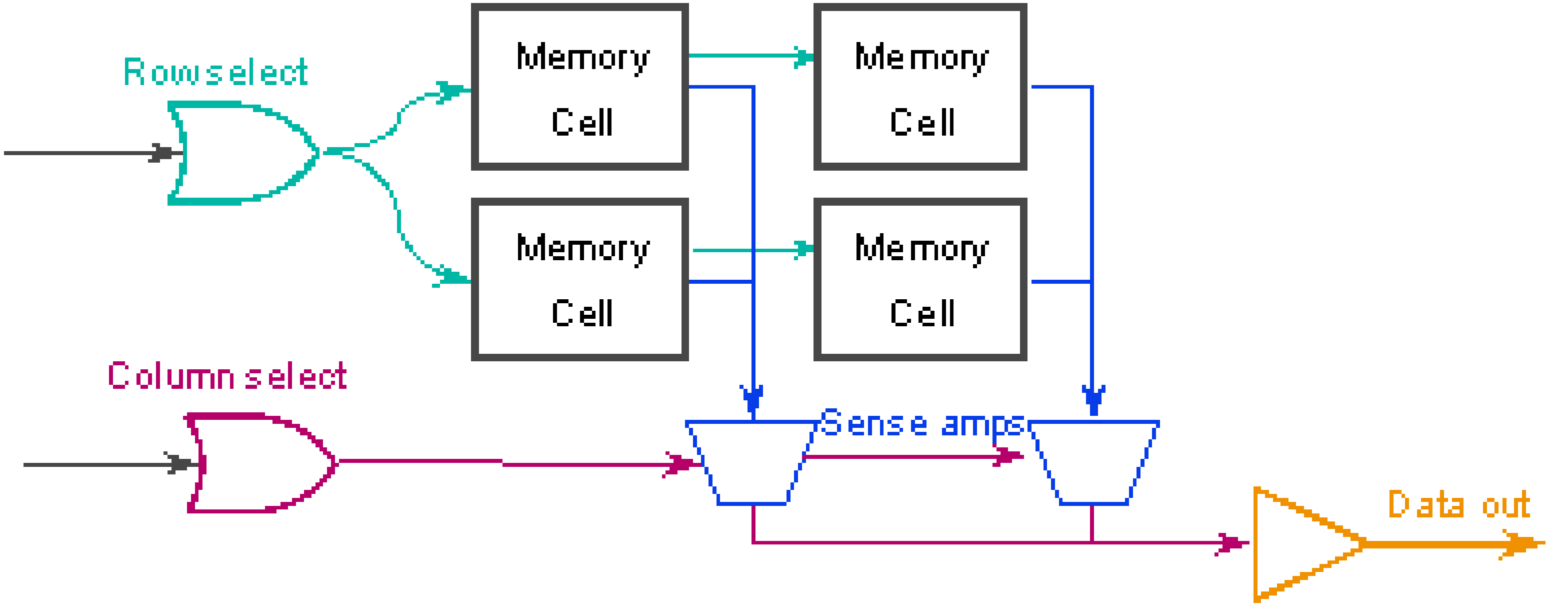

250
200
150
100
50
1,000
3,000
5,000
7,000
9,000

- It is important to note that in the first type of complex system, the general definition does not hold if the parts of the system are identical and if their interconnections are regular.
- A memory chip , has the greatest density of transistors of any integrated circuit, however it is not a complex system. The diagram gives a simple model for the behavior of a dynamic memory device.



The actual design is much more complex, with a large number of carefully interconnected and timed clocks. These clocks interrelate the behavior of the components far more what is indicated.

DRAM



Basic model is simple, but inaccurate

100-200 clocks used for timing:

- Processing and temperature skews
- Allow more time for sense ⇒ squeeze everything before and after

MEASURING COMPLEXITY

- How do we know if the system we are looking at is a complex system?

- How complex is the system?

- In general, the goal is to be able to compare two systems and say how complex they are with respect to each other,, i.e. that system A is more complex than system B, with some sort of general numerical rating scale.

> “My own suspicion is that the universe is not only stranger than we suppose, but stranger than we can suppose.”
> - John Haldane

POSSIBLE APPROACHES

- Human Observation and Rating (subjective)
- Number of parts or Distinct Elements (what is considered a distinct element?)
- Dimensions (how would that be measured?)
- Number of parameters controlling the system (what amount of control?)
- The Minimal Description (in what language?)
- Information content (what would be considered information, and how would it be measured?)
- Minimal generator/ constructer (what machines/ methods would be used?)
- Minimum construction time/ energy (how would evolution count?)

- Most (if not all) of these measures will actually be measures associated with a model of a phenomenon.
- Two observers (of the same phenomenon?) may develop or use very different models, and thus disagree in their assessments of the complexity.
- For example counting the number of parts is likely to depend on the scale at which the phenomenon is viewed.
  - For example: counting atoms is different from counting molecules, cells, organs, etc.

INFORMATION THEORY

- It’s difficult to expect to come up with a single universal measure of complexity. The best we are likely to have is a measuring system useful by a particular observer, in a particular context, for a particular purpose.
- One such system is Information Theory, which reduces any event to the probability of it’s occurrence, and gives a measure of the *information* we get from it.

“Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.”
- John von Neumann (1903-1957)

SOME PROBABILITY

- A frequentist version of probability in which, we assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events, no two of which can occur simultaneously, and the events occur with frequencies n, we say that the probability of event xi is given by:

$$P(x_i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$$

- Which has the useful property:

$$\sum_{i=1}^{N} P(x_i) = 1$$

- In an observer relative version of probability in this version, we take a statement of probability to be an assertion about the belief that a specific observer has of the occurrence of a specific event.
- Note that in this version of probability, it is possible that two different observers may assign different probabilities to the same event.
- Furthermore, the probability of an event, for me, is likely to change as I learn more about the event, or the context of the event.
- In some (possibly many) cases, we may be able to find a reasonable correspondence between these two views of probability. In particular, we may sometimes be able to understand the observer relative version of the probability of an event to be an approximation to the frequentist version.

“The laws of probability, so true in general, so fallacious in particular .”
Edward Gibbon

- Probability basics: Where *a* and *b* are events

$P(\text{not } a) = 1 - a$

$P(a \text{ or } b) = P(a) + P(b) - P(a \text{ and } b)$

$P(a \text{ and } b) = P(a, b)$

If $P(a, b)=0$, a and b are mutually exclusive.

- $P(a|b)$ is the probability of a given that we know b. The joint probability is:

$P(a, b) = P(a|b)P(b)$

- And since $P(a,b)=P(b,a)$ we have Bayes' theorem:

$P(a|b)P(b) = P(b|a)P(a)$

- So: $P(a, b) = P(a|b)\ P(b) = P(a)P(b)$ This is also taken as the definition of independence.

DEVELOPING A USABLE MEASURE

- We would like to develop a usable measure of the information we get from observing the occurrence of an event having probability p . Our first reduction will be to ignore any particular features of the event, and only observe whether or not it happened.
- Thus we will think of an event as the observance of a symbol whose probability of occurring is p, and so we will be defining the information in terms of the probability p.
- This particular approach is *axiomatic*, and we can apply this system in any context in which we have available a set of non-negative real numbers

axiom
ˈaksɪəm/
noun
noun: **axiom**; plural noun: **axioms**
a statement or proposition which is regarded as being established, accepted, or self-evidently true.

THE AXIOMS:

1. Information is a non-negative quantity: $I(p) \geq 0$.

2. If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$.

3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two information(s): $I(p1*p2)=I(p1)+I(p2)$

4. We will want our information measure to be a continuous (and, in fact, monotonic) function of the probability (slight changes in probability should result in slight changes in information).

- From this, we get:

1. $I(p2) = I(p*p) = I(p)+I(p) = 2*I(p)$

2. Thus, further, $I(p^n) = n * I(p)$ (by induction . . .)

3. $I(p) = I((p^{1/m})^m) = m * I(p^{1/m})$, so $I(p^{1/m}) = 1/m * I(P\ )$ and thus in general $I(p^{n/m}) = n/m * I(p)$

4. And thus, by continuity, we get, for $0<p\leq 1$, and $a>0$, a, real number:

   $I(p^a) = a * I(p)$

From this, we can see:

$$\mathbf{I(p) = -\log_b(p) = \log_b(1/p)}$$

- Summarizing: from the four properties,

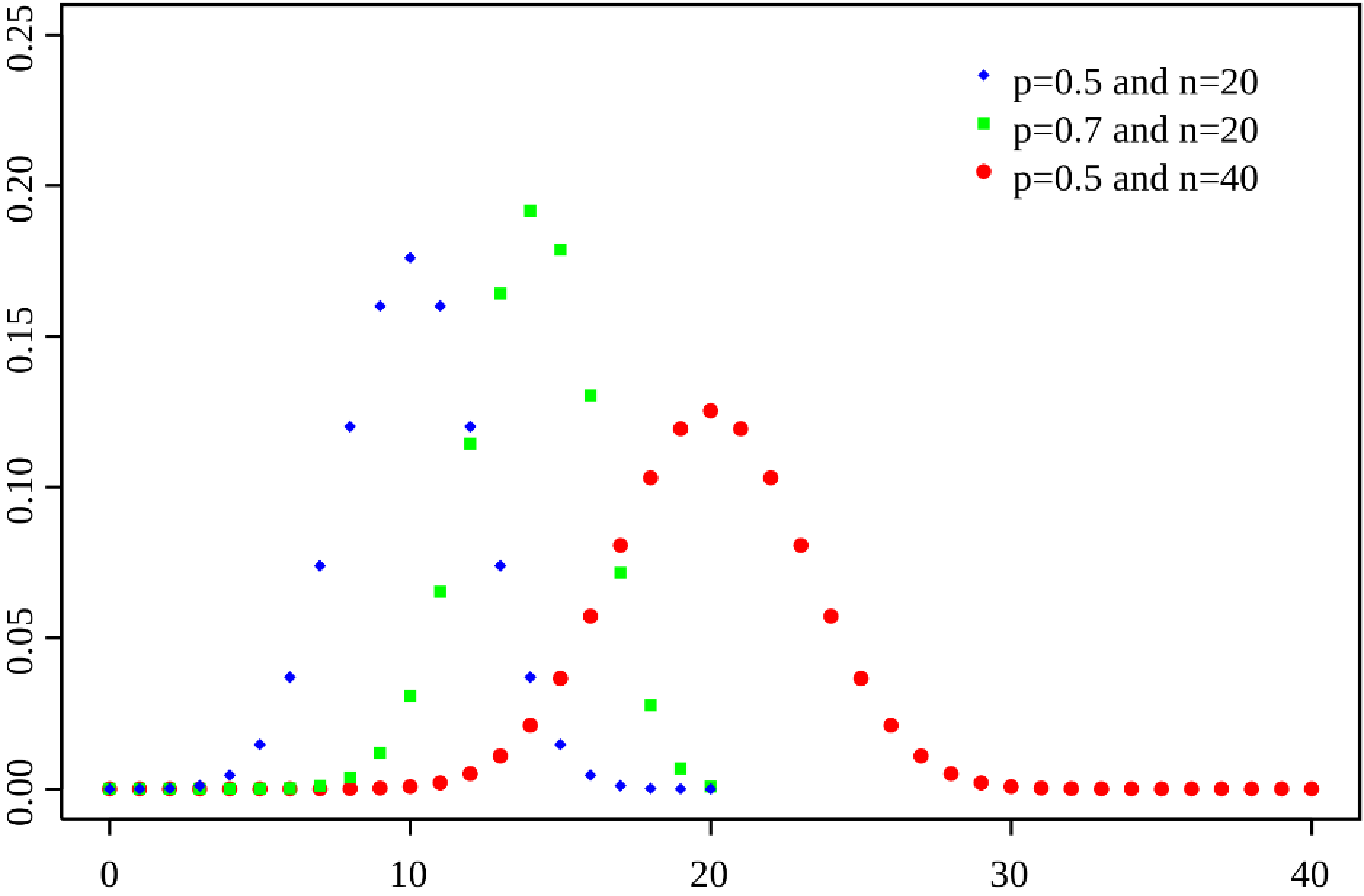
1. 1. I(p) ≥ 0
2. I(p1 ∗ p2) = I(p1) + I(p2)
3. I(p) is monotonic and continuous in p 4.
4. I(1) = 0

We can derive that

$$I(p) = \log_b(1/p) = -\log_b(p)$$

for some positive constant b. The base b determines the units we are using.

Binomial Distribution/
Bernoulli Distribution


p=0.5 and n=20
p=0.7 and n=20
p=0.5 and n=40
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40

- Thus, using different bases for the logarithm results in information measures which are just constant multiples of each other, corresponding with measurements in different units:
  1. log2 units are bits (from ’binary’) –this is what we usually work with
  2. log3 units are trits(from ’trinary’)
  3. loge units are nats (from ’natural logarithm’) (We’ll use ln(x) for loge(x))
  4. log10 units are Hartleys, after an early worker in the field.
- We can change the units by changing the base, using the formulas, for b1, b2, x > 0,

  $x = b_1^{\log b1(x)}$ (x=x) and therefore

  $\log b_2 (x) = \log b_2 (b_1^{\log b1(x)}) = (\log b\ _2(b_1))(\log b_1 (x))$

- For example, flipping a fair coin once will give us events *h* and *t* each with probability 1/2, and thus a single flip of a coin gives us $-\log_2(1/2) = 1$ bit of information (whether it comes up h or t).
- Flipping a fair coin n times (or, equivalently, flipping n fair coins) gives us

$-\log_2((1/2)^n) = \log_2(2^n) = n * \log_2(2) = n$ bits of information.

- So, for 25 flips you could get:

  hthhtththhhthttththhhthtt

- Using 1 for h and 0 for t, the 25 bits

  1011001011101000101110100.

BASIC ENTROPY THEORY

- Entropy can be seen as the randomness of a possible event.
- It's a weighted average of the information carried by an event, or would be carried by the event and can also be the encoding length.

Suppose we had a source emitting *n* symbols $\{a_1,a_2,a_3,a_4\ldots a_n\}$ and with probabilities $\{p_1,p_2,p_3,p_4\ldots p_n\}$ respectively, and that the symbols were being emitted independently, to eliminate conditional probability.

What would be the average information we would get from each symbol in the stream?

- If we observe a_i, we will be getting $\log(1/p_i)$ information from that single observation.
- So if we have N observations in the long run, we will see (approximately) $N*p_i$ occurrences of ai. So in the frequentist sense, that's saying the probability of seeing ai is p_i .
- So from N independent observations we get our information as:

$$I = \sum_{i=1}^{n} (N * p_i) * \log(1/p_i).$$

- But since we want the average information, we divide through by N:

$$\begin{aligned} I/N &= (1/N) \sum_{i=1}^{n} (N * p_i) * \log(1/p_i) \\ &= \sum_{i=1}^{n} p_i * \log(1/p_i) \end{aligned}$$

- This is a fundamental definition of entropy, credited mostly to Shannon, in 1948 (in the seminal papers of the information theory field).
- We defined information strictly in terms of probabilities of event, therefore, if we had a probability distribution,

P={p_1,p_2,p_3,p_4…p_n} we define the entropy of the discrete probability distribution as:

$$H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i).$$

- The generalization is that for continuous instead of discrete probability, the entropy is given as the integration of the function of probability. That is let the probability of X be P(x).
- Therefore entropy for continuous probability is defined as:

$$H(P) = \int P(x) * \log(1/P(x))dx.$$

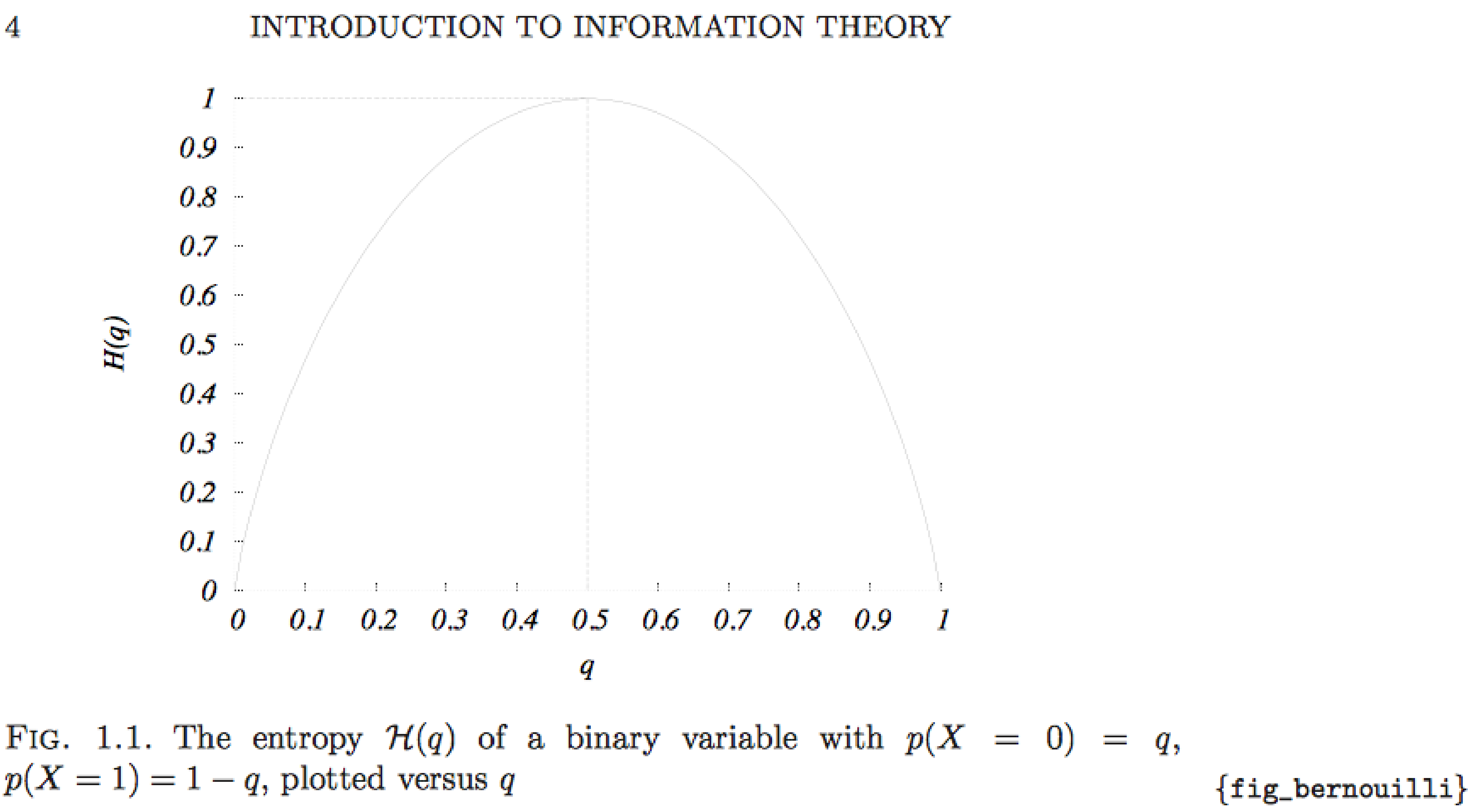
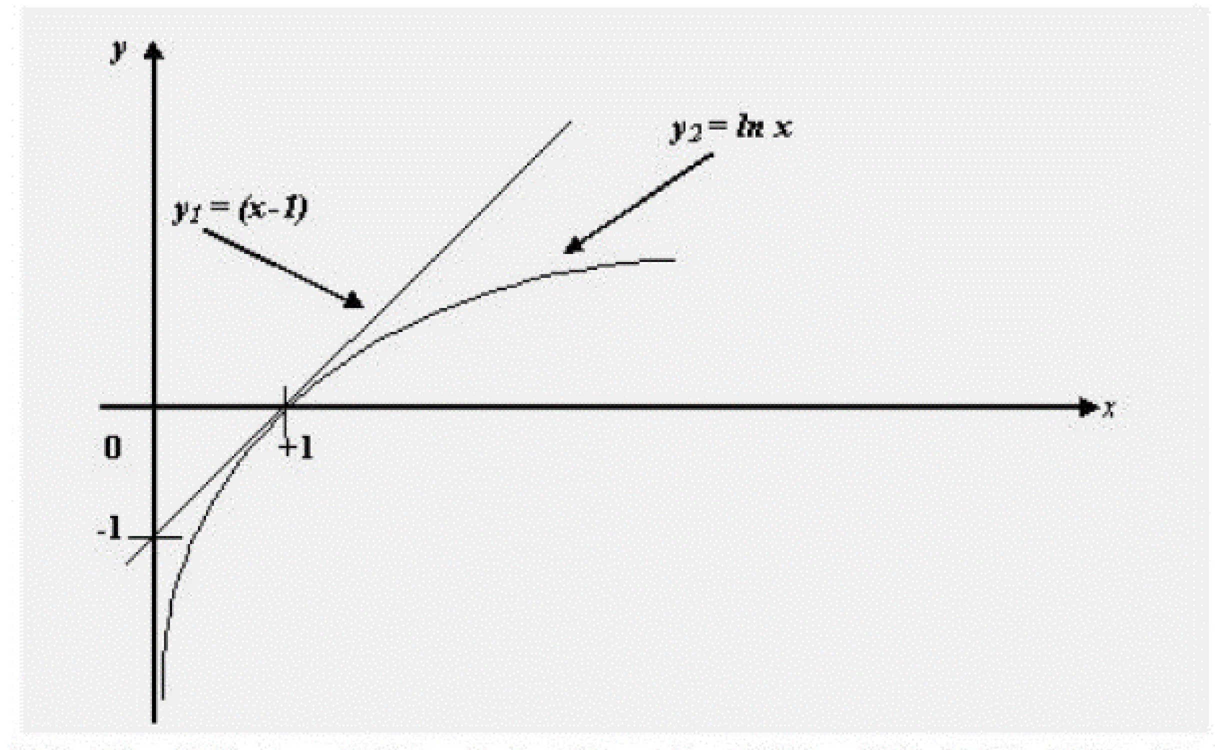


FIG. 1.1. The entropy $\mathcal{H}(q)$ of a binary variable with $p(X = 0) = q$, $p(X = 1) = 1 - q$, plotted versus q {fig_bernouilli}

THE GIBBS' INEQUALITY

- If we look at a continuous probability I(x), when written in terms of information using base e, it becomes ln(x)
- The derivative of ln(x) is 1/x, so we can find the tangent to the curve when x=1 is the line y=x-1
- we also know that the curve of ln(x) is concave down increasing, so we know that x>0, $\ln(x) \leq x-1$ equal only when x=1



- For two probability distributions:
- $P=\{p_1,p_2,p_3,p_4\ldots p_n\}$ and
- $Q=\{q_1,q_2,q_3,q_4\ldots q_n\}$

Where p_i, $q_i \geq 0$ and the sum to i of p_i and q_i is 1 ($\sum_i p_i = \sum_i q_i = 1,$

So, taking x as p_i / q_i in the inequality :

$$\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right) \leq \sum_{i=1}^{n} p_i\left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n}(q_i - p_i)$$

$$= \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0,$$

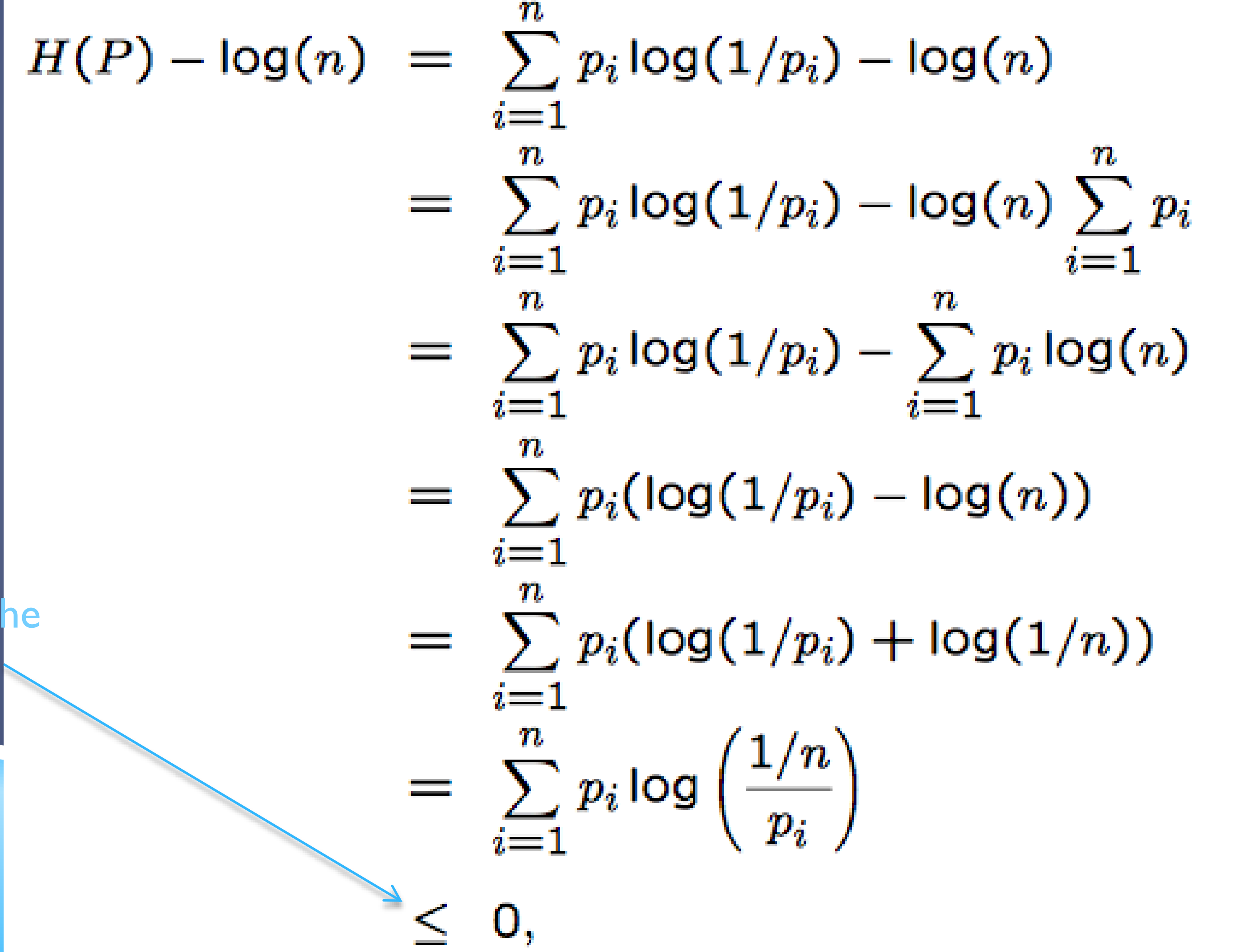
From:

$$= \sum_{i=1}^{n} p_i * \log(1/p_i)$$

We use base e here as that is what was used to find the inequality, but it's clear that it holds for all bases.

FINDING A PROBABILITY DISTRIBUTION WHICH MAXIMIZES THE ENTROPY FUNCTION

- If we say $P=\{p_1,p_2,p_3,p_4\ldots p_n\}$ is a probability distribution, then:



$$\begin{aligned}
H(P) - \log(n) &= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \sum_{i=1}^{n} p_i \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \sum_{i=1}^{n} p_i \log(n) \\
&= \sum_{i=1}^{n} p_i(\log(1/p_i) - \log(n)) \\
&= \sum_{i=1}^{n} p_i(\log(1/p_i) + \log(1/n)) \\
&= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right) \\
&\leq 0,
\end{aligned}$$

With equality only when p=1/n for all values of i

Applying the
nequality)

- This means that:

$$0 \leq H(P) \leq \log(n)$$

- We have H(P) = 0 when exactly one of the pi's is one and all the rest are zero.
- We have H(P) = log(n) only when all of the events have the same probability 1/n.
- That is, the maximum of the entropy function is the log() of the number of possible events, and occurs when all the events are equally likely.

HOW MUCH INFORMATION IS THERE IN A GRADE?

- Firstly, to maximize the entropy (information) of a simple pass fail distribution, on average about half the class should pass, since you want an equal probability of 1/n.
- The maximum information the student gets from a grade will be:
- Pass/Fail : 1 bit. [$\log_2(2)=1$]
- A, B, C, D, F : 2.3 bits. [$\log_2(5)= 2.32192809489$]
- A,A-,B+,. . .,D-,F: 3.6bits. [$\log_2(12)= 3.58496250072$]

NOTES:

- First, the definitions of information and entropy may not match with some other uses of the terms.
- For example, if we know that a source will, with equal probability, transmit either the complete text of Hamlet or the complete text of Macbeth (and nothing else), then receiving the complete text of Hamlet provides us with precisely 1 bit of information.
- Suppose a book contains ASCII characters. If the book is to provide us with information at the maximum rate, then each ASCII character will occur with equal probability – it will be a random sequence of characters.

- Second, it is important to recognize that our definitions of information and entropy depend only on the probability distribution. In general, it won't make sense for us to talk about the information or the entropy of a source without specifying the probability distribution.
- Two different observers of the same data stream have different models of the source, and thus associate different probability distributions to the source. The two observers will then assign different values to the information and entropy associated with the source.
- This observation (almost :-) accords with our intuition: two people listening to the same lecture can get very different information from the lecture.

THANK YOU