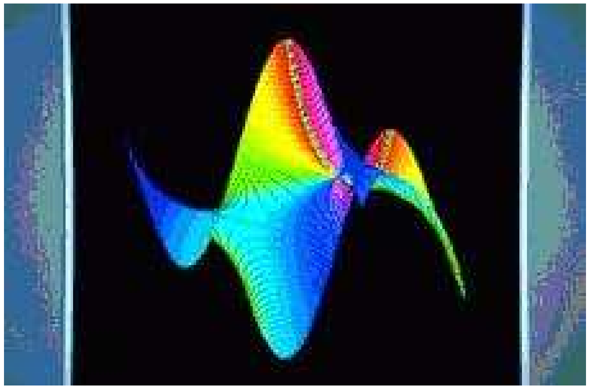
What is Information?*

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

April 30, 2008

AofA and **IT** logos



INRIA 2008

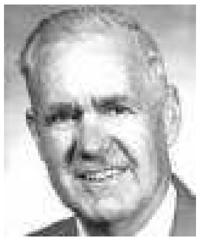
*Participants of Information Beyond Shannon, Orlando, 2005, and J. Konorski, Gdansk, Poland.

Outline

1. Standing on the Shoulders of Giants . . .
2. What is Information?
3. Shannon Information
   - Beyond Shannon
   - Temporal and Darwin Channels
4. Physics of Information
   - Shannon vs Boltzmann
   - Maxwell's Demon, Szilard's Engine, and Landauer's Principle
5. Ubiquitous Information (Biology, Chemistry, Physics)
6. Today's Challenges
7. Science of Information

Standing on the Shoulders of Giants . . .

C. F. Von Weizsäcker:



"Information is only that which produces information" (relativity).
"Information is only that which is understood" (rationality)
"Information has no absolute meaning".

R. Feynman:

". . . Information is as much a property of your own knowledge as anything in the message.
. . . Information is not simply a physical property of a message: it is a property of the message and your knowledge about it."

J. Wheeler:
"It from Bit". (Information is physical.)

C. Shannon:
"These semantic aspects of communication are irrelevant . . ."

Structural and Biological Information

F. Brooks, jr. (JACM, 50, 2003, "Three Great Challenges for . . . CS "):

"Shannon and Weaver performed an inestimable service by giving us a definition of Information and a metric for for Information as communicated from place to place. We have **no theory** however that gives us a metric for the Information embodied in **structure** . . . this is the most fundamental gap in the theoretical underpinning of Information and computer science. . . . A young information theory scholar willing to spend years on a deeply fundamental problem need look no further."

M. Eigen

"The differentiable characteristic of the living systems is **Information**. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability . . . . Information theory, pioneered by **Claude Shannon**, cannot answer this question . . . in principle, the answer was formulated 130 years ago by **Charles Darwin**.

What is then Information?

Information has the flavor of:

relativity (depends on the activity undertaken),

rationality (depends on the recipient's knowledge),

timeliness (temporal structure),

space (spatial structure).

Informally Speaking: A piece of data carries **information** if it can impact a **recipient's ability** to achieve the **objective** of some **activity** within a given **context**.

Using the event-driven paradigm, we may formally define:

Definition 1. *The* **amount of information** *(in a faultless scenario)* info*(E) carried by the event E in the context C as measured for a system with the rules of conduct R is*

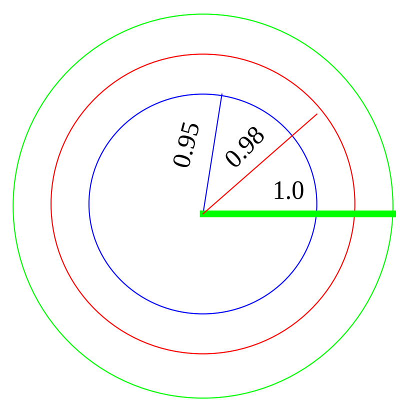
$$\mathrm{info}_{R,C}(E) = \mathbf{cost}[\mathrm{objective}_R(C(E)), \mathrm{objective}_R(C(E)+E)]$$

where the **cost** *(weight, distance) is taken according to the ordering of points in the space of objectives.*

Example: Decimal Representation

Example 1: In a decimal representation of π, the objective is to learn the number π and P is to compute successive digits approximating π.

Imagine we are drawing circles of circumferences, i.e., $3, 3.1, 3.14, 3.141$ etc., and measure the respective diameters i.e., $.9549, .9868, .9995, .9998$, which asymptote to the ideal 1.



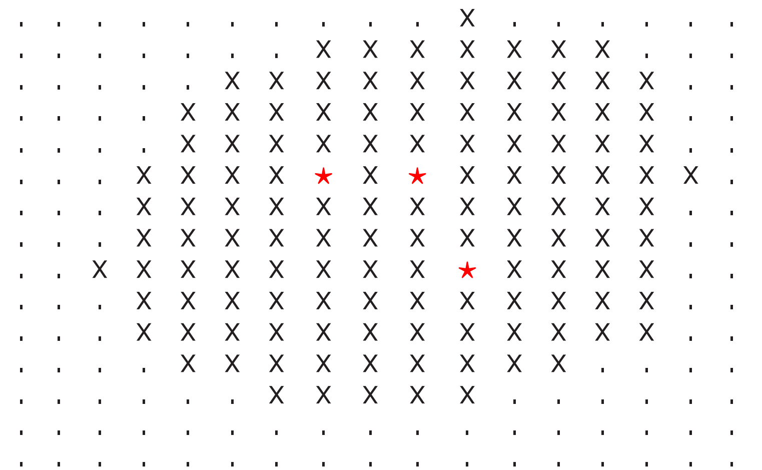
info = information is the the difference between successive deviations from the ideal 1.

For example:

- event "3" carries $(1-0)-(1-.9549) = .9549$,
- event "1" carries $(1-.9549)-(1-.9868) = .0319$,
- event "4" carries $(1-.9995)-(1-.9868) = .0127$, etc.

Example: Distributed Information

1. Example 2: In an N-threshold secret sharing scheme, N subkeys of the decryption key roam among $A \times A$ stations.

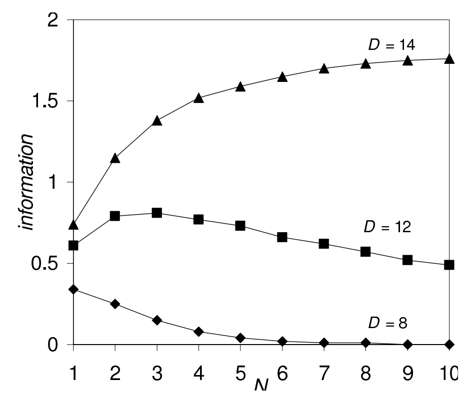


2. By protocol P a station has access:

- only it sees all N subkeys.
- it is within a distance D from all subkeys.

3. Assume that the larger N, the more valuable the secrets. We define the amount of information as

info= $N \times \{\text{\# of stations having access}\}$.



Outline Update

1. Standing on the Shoulders of Giants . . .
2. What is Information?
3. Shannon Information
4. Physics of Information
5. Ubiquitous Information
6. Today's Challenges

Shannon Information . . .

In 1948 C. Shannon created a powerful and beautiful **theory of information** that served as the backbone to a now classical paradigm of digital communication.

In our setting, Shannon defined:

objective: statistical ignorance of the recipient;
statistical uncertainty of the recipient.

cost: # binary decisions to describe E;
= $-\log P(E)$; $P(E)$ being the probability of E.

Context: the semantics of data is irrelevant . . .

Self-information for E_i: $\text{info}(E_i) = -\log P(E_i)$.
Average information: $H(P) = -\sum_i P(E_i) \log P(E_i)$
Entropy of $X = \{E_1, \ldots\}$: $H(X) = -\sum_i P(E_i) \log P(E_i)$
Mutual Information: $I(X;Y) = H(Y) - H(Y|X)$, (faulty channel).

Shannon's statistical information tells us how much a recipient of data can reduce their statistical uncertainty by observing data.

Shannon's information is not absolute information since $P(E_i)$ (prior knowledge) is a subjective property of the recipient.

Shortest Description, Complexity

Example: X can take eight values with probabilities:

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right).$$

Assign to them the following code:

$$0, 10, 110, 1110, 111100, 111101, 111110, 111111,$$

The entropy X is

$$H(X) = 2 \text{ bits}.$$

The shortest description (on average) is 2 bits.

In general, if X is a (random) sequence with entropy $H(X)$ and average code length $L(X)$, then

$$H(X) \leq L(X) \leq H(X) + 1.$$

Complexity vs Description vs Entropy

The more complex X is, the longer its description is, and the bigger the entropy is.

Three Jewels of Shannon

Theorem 1. (**Shannon 1948; Lossless Data Compression**).

compression bit rate $\geq$ source entropy $H(X)$.

(There exists a codebook of size 2^{nR} of universal codes of length n with

$$R > H(X)$$

and probability of error smaller than any $\varepsilon > 0$.)

Theorem 2. (**Shannon 1948; Channel Coding**)

In Shannon's words:

It is possible to send information at the capacity through the channel with as small a frequency of errors as desired by proper (**long**) encoding. This statement is not true for any rate greater than the capacity.

(The maximum codebook size $N(n, \varepsilon)$ for codelength n and error probability ε is asymptotically equal to: $N(n, \varepsilon) \sim 2^{nC}$.)

Theorem 3. (**Shannon 1948; Lossy Data Compression**).

For distortion level D:

lossy bit rate $\geq$ rate distortion function $R(D)$.

Rissanen's MDL Principle

1. Objective(P, C) may include the cost of the very recognition and interpretation of C.

2. In 1978 Rissanen introduced the Minimum Description Length (MDL) principle (Occam's Razor) postulating that the best hypothesis is the one with the shortest description.

3. Universal data compression is used to realize MDL.

4. Normalized maximum likelihood (NML) code: Let $\mathcal{M}_k = \{Q_\theta : \theta \in \Theta\}$ and let $\hat{\theta}$ minimize $-\log Q_\theta(x)$. The minimax regret is

$$r_n^*(\mathcal{M}) = \min_Q \max_x \left[\log \frac{Q_{\hat{\theta}}(x)}{Q_\theta(x)}\right] = \log \sum_x Q_{\hat{\theta}}(x) = \log \sum_x \sup_\theta Q_\theta(x).$$

Rissanen proved for memoryless and Markov sources:

$$r_n^*(\mathcal{M}_k) = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_\theta \sqrt{|I(\theta)|} d\theta + o(1).$$

5. Why to restrict analysis to prefix codes? Fundamental lower bound? For one-to-one codes (cf. W.S., ISIT, 2005).

$$\text{redundancy} = -\frac{1}{2} \log n + O(1)?$$

Beyond Shannon

Participants of the **2005 Information Beyond Shannon** workshop realize:

Delay: In networks, delay incurred is a issue not yet addressed in information theory (e.g., complete information arriving late maybe useless).

Space: In networks the spatially distributed components raise fundamental issues of limitations in information exchange since the available resources must be shared, allocated and re-used. **Information** is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective.

Structure: We still lack measures and meters to define and appraise the amount of information embodied in structure and organization.

Semantics. In many scientific contexts, one is interested in signals, without knowing precisely what these signals represent. What is semantic information and how to characterize it? How much more semantic information is there when when compared with its syntactic information?

Limited Computational Resources: In many scenarios, information is limited by available computational resources (e.g., cell phone, living cell).

Physics of Information: Information is physical (J. Wheeler).

Some Things to Think About . . .

Here is a short list of "toy problems" to think about:

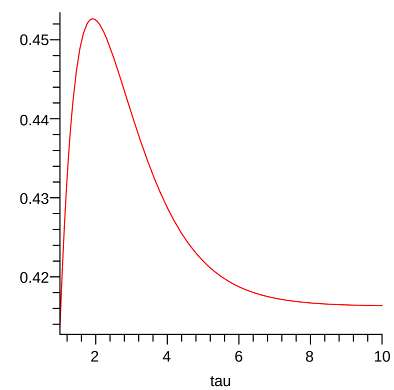
- Temporal Capacity (e.g., assign transmission time to each symbol or a block of symbols).
- Spatial Capacity (e.g., destination may be in different locations).
- Darwin Channel that models flow of genetic information (e.g., a combination of a deletion/insertion channel and constrained channel).
- Distributed Information (information here/local and there/distributed is not the same?)
- Speed of Information (how fast information can be spread out?)
- Entropy of a structure (e.g., graph entropy)
- Representation-invariant measure of information (Shannon, 1953).

Temporal Capacity

1. Binary symmetric channel (BSC): each bit incurs a delay.
2. Delay T has known probability distribution: $F(t) = P(T < t)$.
If a bit arrives after a given deadline τ, it is dropped.
3. The longer it takes to send a bit, the lower the probability of a success, which we denote by $\Phi(\varepsilon, t)$ for $t < \tau$ (e.g., $\Phi(\varepsilon, t) = (1-\varepsilon)^t$).
4. Define $P(x|x) = \int_0^\tau \Phi(\varepsilon, t) dF(t)$: prob. of a successfully transmission:

$$P(y|x) = \begin{cases} \alpha := 1 - F(\tau) & y = \text{erasure} \\ P(x|x) & \text{if } x = y \\ 1 - \alpha - P(x|x) & \text{if } x \neq y. \end{cases}$$

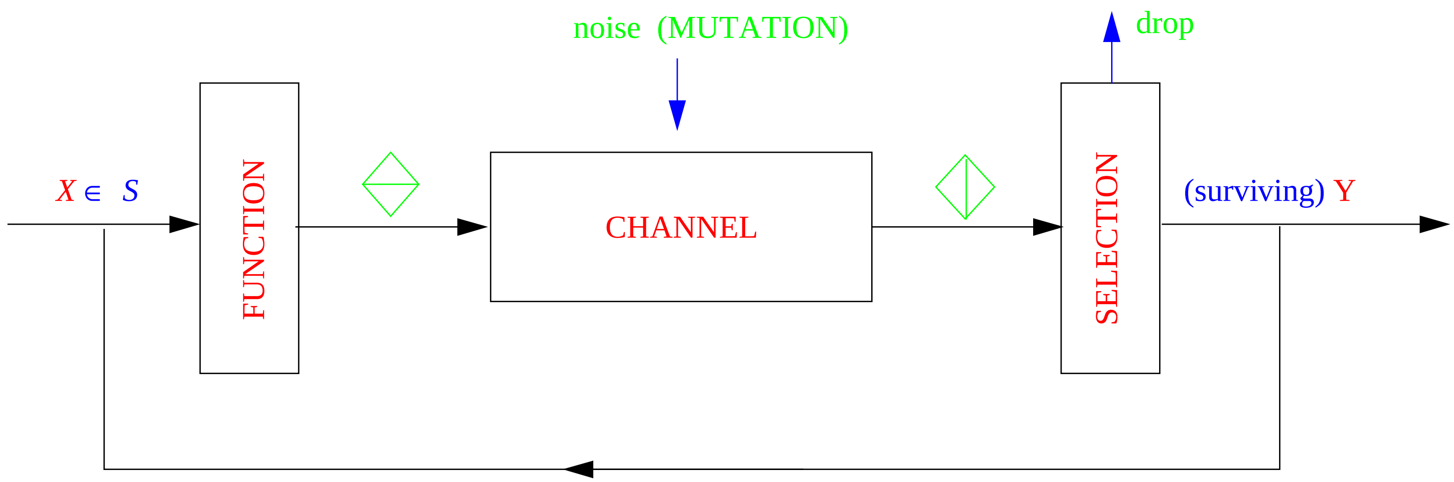
5. Define: $\alpha = 1 - F(\tau)$ and $\rho := \frac{P(x|x)}{(1-\alpha)}$.
Note $C(\tau) := H(Y) - H(Y|X)$, where
$H(Y|X) = H(\alpha) + (1-\alpha)H(\rho)$ and
$H(Y) = H(\alpha) + (1-\alpha)H(p\rho + \bar{p}\bar{\rho})$.
Then:

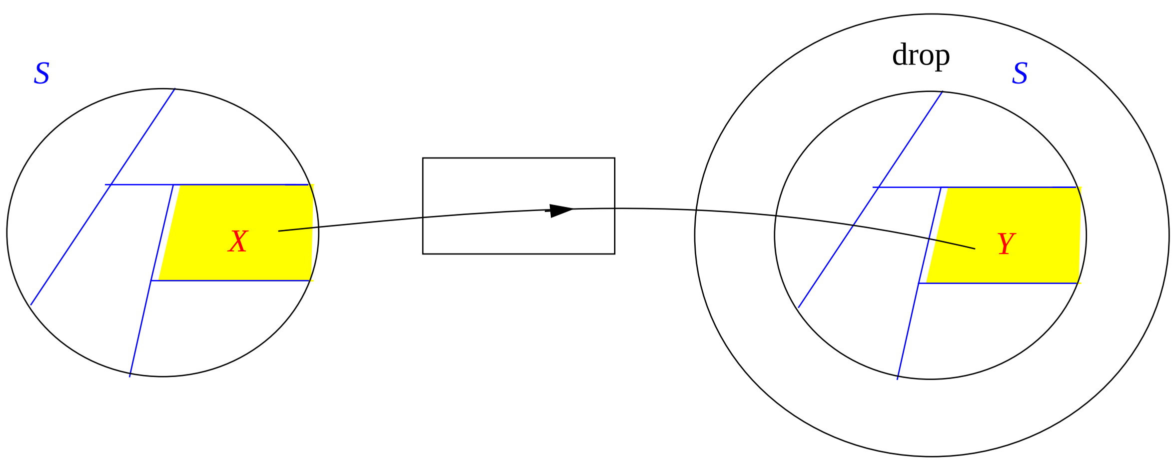


$$C(\tau) = [(1 - P(T > \tau)][1 - H(\rho)].$$

Darwin Channel

To capture sources of variation and natural selection, one is tempted to introduce the so called **Darwin channel** that models the flow of genetic information. (Special case of it is the noisy constrained channel.)



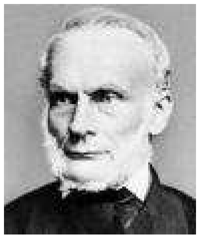


Outline Update

1. Standing on the Shoulders of Giants . . .
2. What is Information?
3. Shannon Information
4. Physics of Information
   - Shannon vs Boltzmann
   - Maxwell's Demon, Szilard's Engine, and Landauer's Principle
5. Ubiquitous Information
6. Today's Challenges

R. Feynmann (*Lectures on Computation*):
. . . information is proportional to the free energy required to reset a "tape" (message) to a fixed state . . .

Clausius and Boltzmann Entropies



R. Clausius in 1850 defined entropy as

$$dS = \frac{dQ}{T}$$

where Q is heat and T temperature.

Boltzmann defined in 1877 statisticall entropy S as

$$S = k \log W$$

where W is the number of molecule macrostates, and k is Boltzmann's constant (to get the correct units).

How do we interpret Boltzmann's entropy?

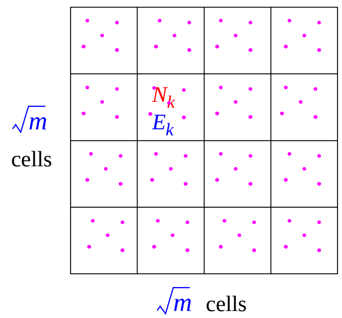
Boltzmann wanted to find out:

How are molecules distributed?

How are Clausius, Shannon and Boltzmann's entropies related (cf. Brillouin, Jaynes, Tribus)?

Boltzmann $\to$ Shannon

Divide space into m cells each containing N_k molecules with energy E_k.
How many configurations, W, are there?



$$W = \frac{N!}{N_1!N_2!\cdots N_m!}$$

$$\text{subject to} \quad N = \sum_{i=1}^{m} N_i, \qquad E = \sum_{i=1}^{m} N_i E_i$$

Boltzmann asked: Which distribution is the most likely to occur?

Boltzmann's answer: the most probable distribution is the one that occurs in the greatest number of ways!
Solving the constrained optimization problem, we find (cf. E. Jaynes)

$$\begin{aligned} \log W &= -N \textstyle\sum_{i=1}^{m} \left(\frac{N_i}{N}\right) \log\left(\frac{N_i}{N}\right) \\ &= N H(P) \qquad \left(\text{for } P_i = \frac{N_i}{N} = \frac{e^{-\beta E_i}}{Z} = \frac{e^{-\beta E_i}}{\sum_i e^{-\beta E_i}}\right), \end{aligned}$$

$\beta = \frac{1}{kT}$ is the Largrange's multiplier, and $Z = \sum_i e^{-\beta E_i}$ is the partition function.

Shannon → Clausius

We start from

$$S = k \log W = -k \sum_i P_i \log P_i$$

where

$$P_i = \frac{N_i}{N} = \frac{e^{-\beta E_i}}{Z}, \quad \beta = \frac{1}{kT}.$$

But

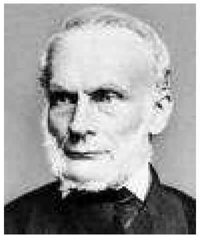
$$dS = -k \sum_i (dP_i \log P_i + dP_i) = -k \sum_i dP_i \log P_i$$

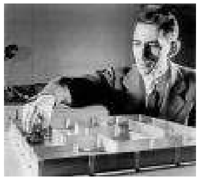
since $d \sum_i P_i = 0$.

Observe that $\log P_i = -\beta E_i - \log Z$, hence

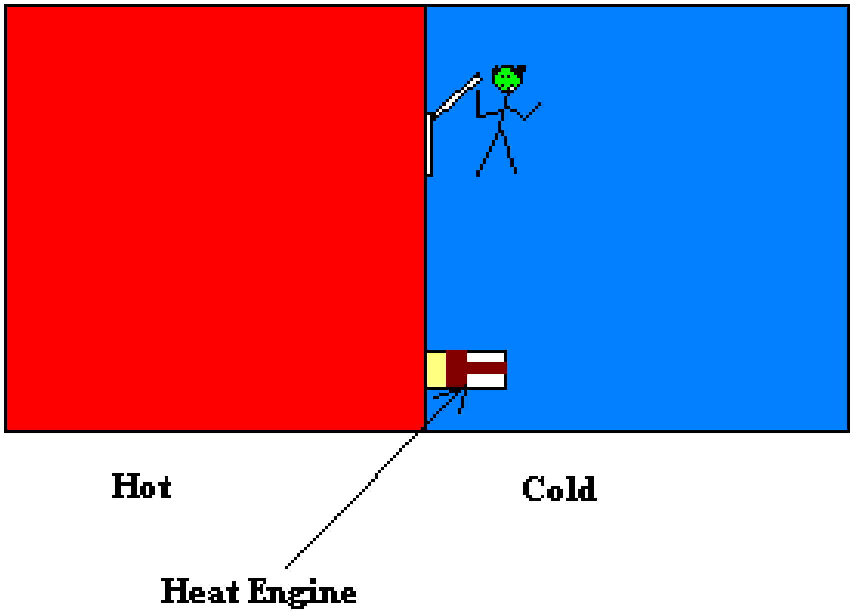
$$dS = k\beta \sum_i E_i dP_i = \frac{dQ}{T}$$

where $dQ = \sum_i E_i dP_i$ represents heat transfered from the outside.





Maxwell's Demon



Second Law of Thermodynamics:

Physics:

$\Delta S \geq 0$

Information Theory (T. Cover)

For Markov source X_n:

(i) with uniform statationary dist.:

$H(X_{n+1}) \geq H(X_n)$;

(ii) stationary Markov source:

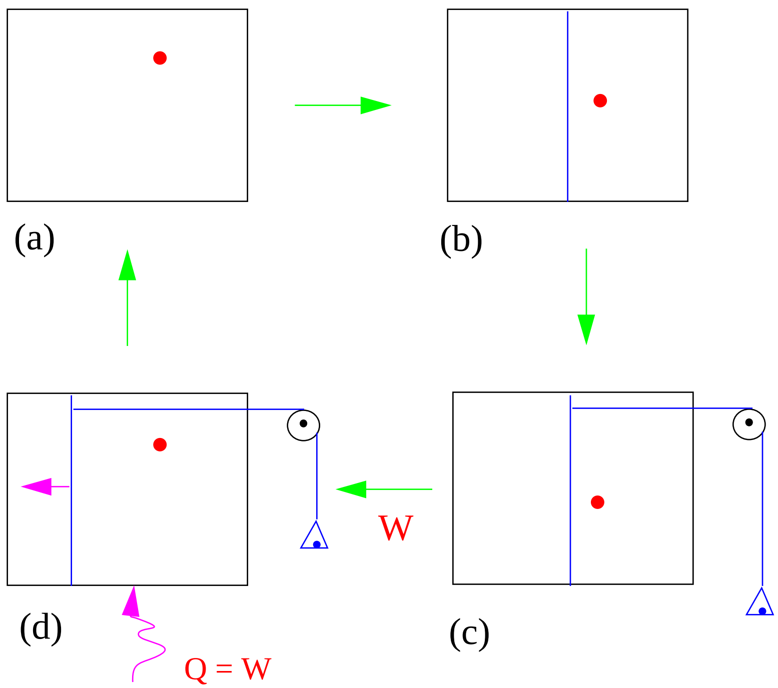
$H(X_{n+1}|X_1) \geq H(X_n|X_1)$.

The second law of thermodynamics:

> the total entropy of any thermodynamically isolated system tends to increase over time, that is, $\Delta S \geq 0$.

Is the Second Law of Thermodynamics violated by Maxwell's Demon?

Szilard's Engine



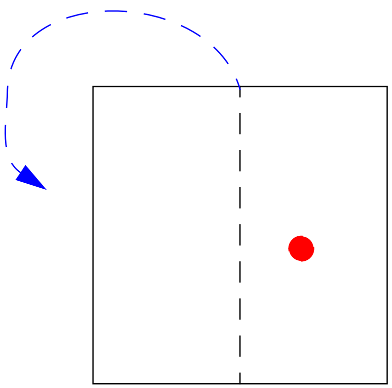
Szilard's Engine: (Acquiring) Information $\Rightarrow$ Energy.

Energy: $E = kNT\ln(V_F/V_I)$.
In Szilard's engine, there is one molecule, hence $E = kT\ln(2)$.

1 bit of **information** $= kT\ln 2$ (joules) of **energy**.

Landauer's Principle: Limits of Computations

Landauer's Principle (1961): Rolf Landauer argued in 1961 that: "any logically irreversible manipulation of information (e.g., erasure of a bit or the merging of two computations) is also physical irreversible and must be accompanied by a corresponding entropy increase . . .".

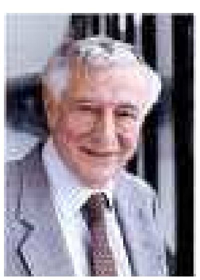


Information erasure $\equiv$ amount of energy to reset a system to zero (asymmetry of resetting). Information is there only only if we randomize the molecule.

By Szilard: Energy equals $T\Delta S = kT\log(V_F/V_I) = kT\log 2$.

By Boltzmann: : Energy equals $T\Delta S = kT\log(W) = kT\log 2$.

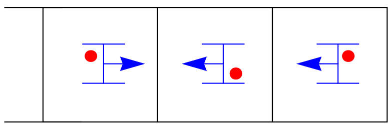
von Neumann-Landauer Bound:

$\text{Irreversible computation} \geq kT\ln 2$ (joules)

Maxwell's demon Explained

R. Feynmann (*Lectures on Computation*):
. . . information is proportional to the free energy required to reset a "tape" (message) to a fixed state . . .



C. H. Bennett:
How much work (fuel) can be extracted from a tape?
Let I be information and N number of bits (molecules):
$\text{work} = (N - I)kT \log 2.$

Maxwell's demon explained:

C.H. Bennett observed that to determine what side of the gate a molecule must be on, the demon must store information about the state of the molecule.

Eventually(!) the demon will run out of information storage space and must begin to erase the information and by Landauer's principle this will increase the entropy of a system.

Bennett's Argument

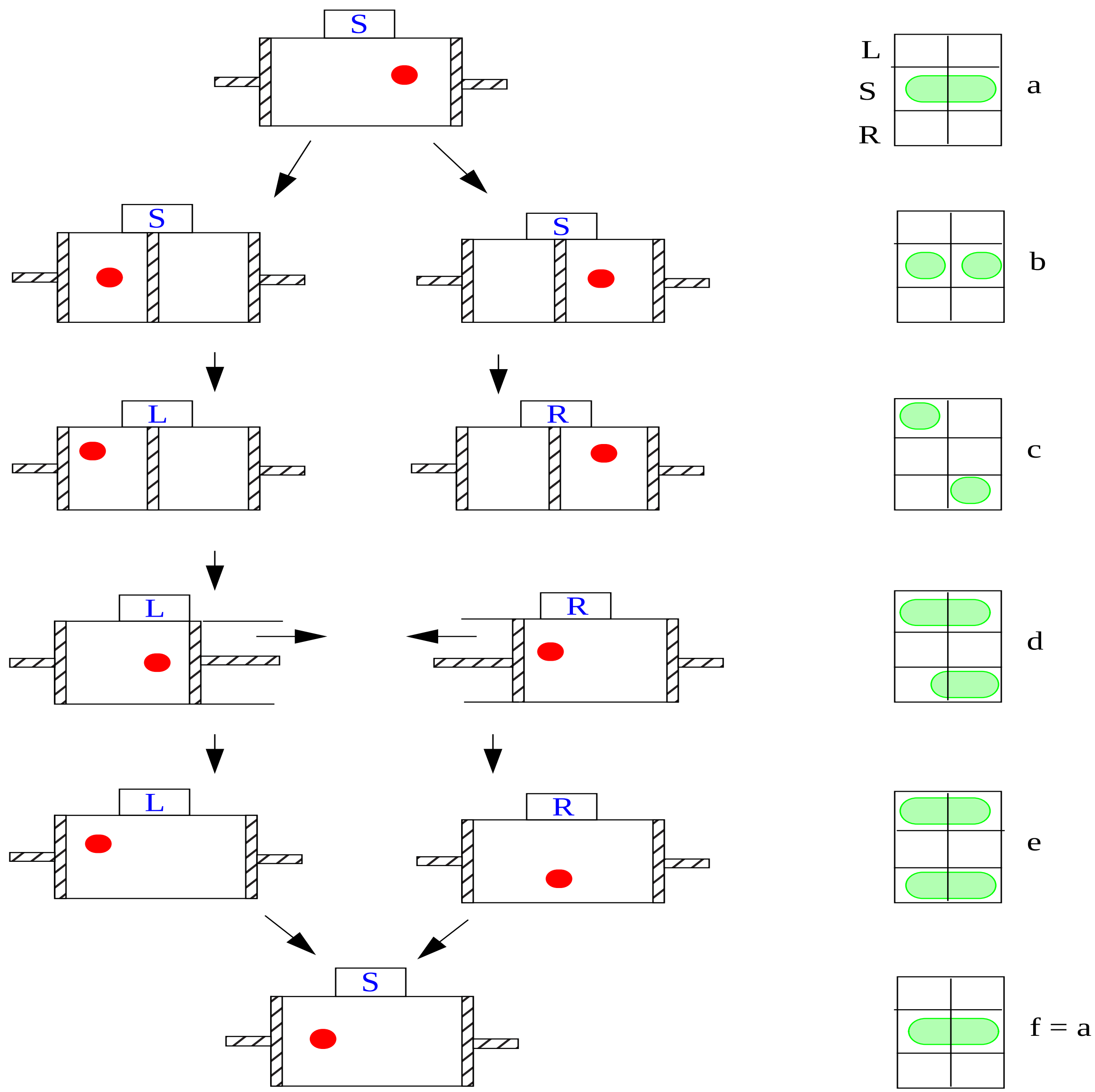


Figure 1: Bennett's argument: erasure of information not measurement is the source of entropy generation.

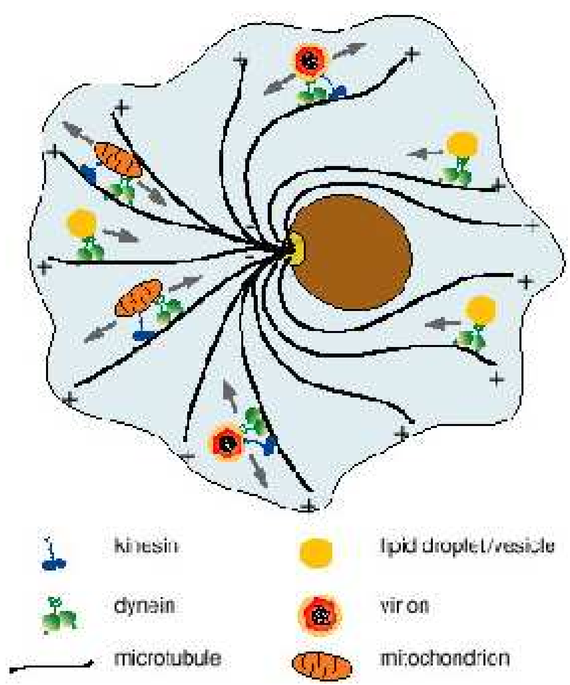
Outline Update

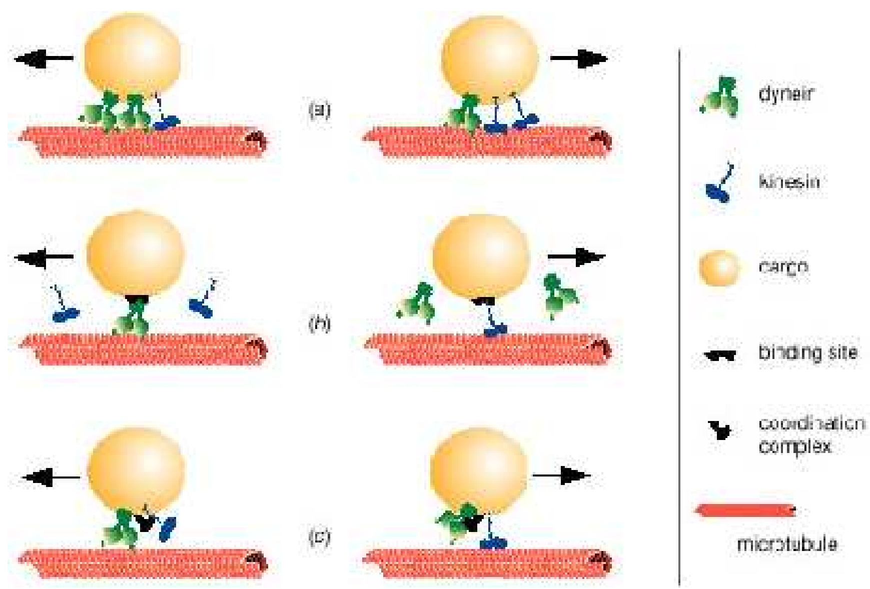
1. Standing on the Shoulders of Giants . . .
2. What is Information?
3. Shannon Information
4. Physics of Information
5. Ubiquitous Information (Biology, Chemistry, Economics, Physics)
6. Today's Challenges

Ubiquitous Information (Biology)

Life is a delicate interplay of energy, entropy, and information; essential functions of living beings correspond to the generation, consumption, processing, preservation, and duplication of **information**.

- How information is generated and transferred through underlying mechanisms of variation and selection (**Darwin channel**).
- How information in biomolecules (sequences and structures) relates to the organization of the cell.
- Whether there are error correcting mechanisms (codes) in biomolecules.
- How organisms survive and thrive in noisy environments.

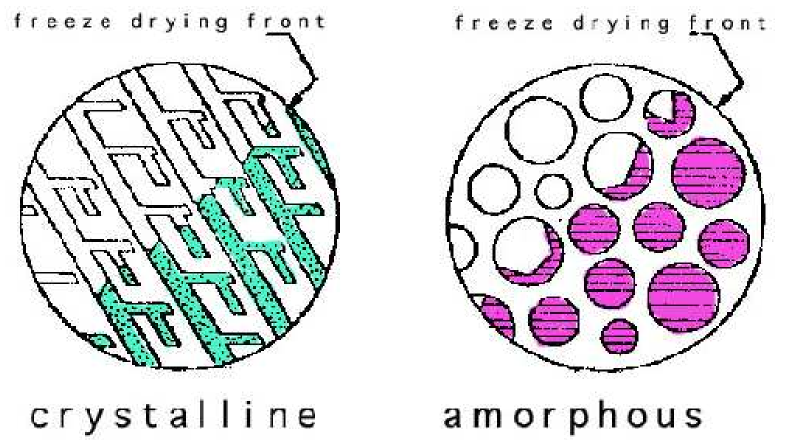




Ubiquitous Information (Chemistry)

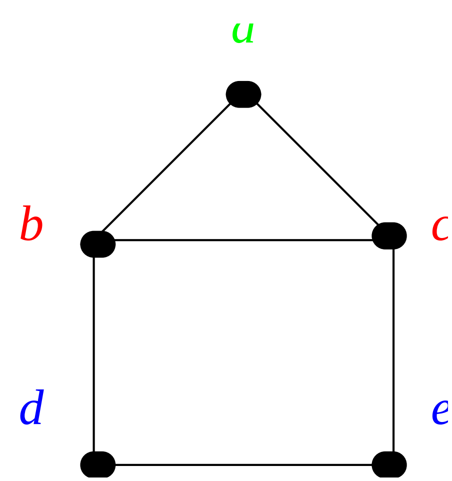
In chemistry information may be manifested in **shapes** and **structures**.

Amorphous solid:
no long-range order.
Crystalline solids:
long-range atomic order.



By how much are amorphous solids more **complex** than crystalline solids?

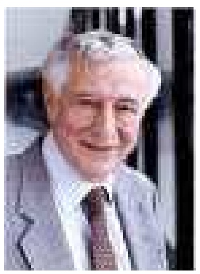
Structural Information
Distinctivness of nodes:
$\{a\}$,
$\{b, c\}$,
$\{d, e\}$.
Group Automorphism and Orbits.



Quantum Information

The laws of nature . . . in quantum theory . . . deal with
. . . our knowledge of the elementary particles.
The concept of objective reality . . . evaporates into . . .
our knowledge of its behavior.

. . . any attempt to measure (that) property
destroys (at least partially) the influence
of earlier knowledge of the system.



. . . the laws of physics are limited
by the range of information processing available.

. . . reality and information are two sides of the same coin,
that is, they are in a deep sense indistinguishable.

Limited Resources Information in Quantum

Quantum physics is a theory of information for systems with limited information resources (Brukner & Zeilinger, 2006).

Brukner and Zeilinger (2001, 2006) postulate:

1. The information content of a quantum system is finite.
   - . . . randomness is a direct consequence of the fact that no enough information is available to pre-define the outcomes of all possibilities.
   - For complementarity the information available suffices to define the outcomes of mutually complementary measurements.
   - Entanglement is a consequence of finite information available to characterize only joint observations (Zeilinger, Nature, 2005).
2. The most elementary system represents the truth value of one proposition.
3. The most elementary system carries 1 bit of information.
4. N elementary systems carry N bits of information.

Shannon Postulates for Entropy

1. H should be continuous in the p_i.

2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three

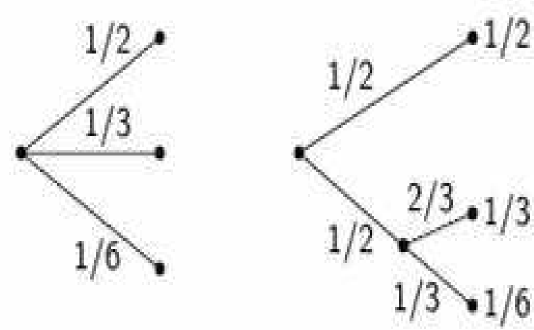


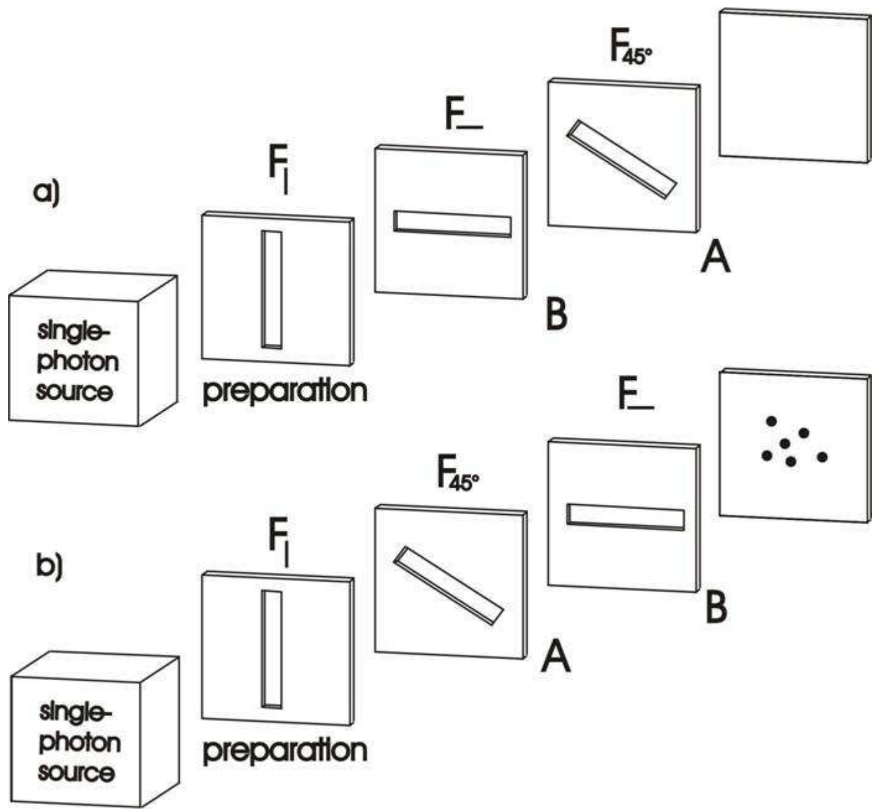
Fig. 6—Decomposition of a choice from three possibilities.

possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}H(\tfrac{2}{3}, \tfrac{1}{3}).$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

Brukner-Zeilinger Experiment



In quantum mechanics events do not necessarily commute, thus

$$H(A,B) \neq H(B,A)\ (!)$$

and potentially

$$H(B) < H(B|A)\ (!).$$

Brukner & Zeilinger suggest to define total information in quantum as

$$I(p_1,\ldots,p_n) = \sum_{i=1}^{n}\left(p_i - \frac{1}{n}\right)^2 = \sum_i p_i^2 - \frac{1}{n} = 2^{-H_2(p)} - \frac{1}{n}$$

so that the sum of individual measures over mutually complementary measurements is invariant under unitary transformations.

Law of Information?

The flow of information about an object into its surrounding is called **decoherence** (increases entanglement with its environment) (H. Zeh, W. Zurek).

Decoherence occurs very, very, very fast, in $10^{-10} - 10^{-20}$ seconds.

The essential difference between microscopic world (quantum) and macroscopic world is **decoherence**.

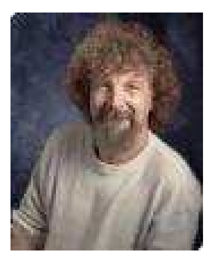
Entropy and decoherence are related, but while entropy operates on a time scale of microseconds, decoherence works a billion times faster.

A new law of Information(?):

> Information can be neither created nor destroyed.

or perhaps

> stored information of any "isolated system" tends to dissipate.



Today's Challenges

- We still lack measures and meters to define and appraise the amount of structure and organization embodied in artifacts and natural objects.

- Information accumulates at a rate faster than it can be sifted through, so that the bottleneck, traditionally represented by the medium, is drifting towards the receiving end of the channel.

- Timeliness, space and control are important dimensions of Information. Time and space varying situations are rarely studied in Shannon Information Theory.

- In a growing number of situations, the overhead in accessing Information makes information itself practically unattainable or obsolete.

- Microscopic systems do not seem to obey Shannon's postulates of Information. In the quantum world and on the level of living cells, traditional Information often fails to accurately describe reality.

- What is the impact of rational/noncooperative behavior on information? What is the relation between value information and information?

Science of Information

Institute for Science of Information

At Purdue we initiated the

Institute for Science of Information

integrating research and teaching activities aimed at investigating the role of **information** from various viewpoints: from the fundamental theoretical underpinnings of information to the science and engineering of novel information substrates, biological pathways, communication networks, economics, and complex social systems.

The specific means and goals for the Center are:

- continue the Prestige Science Lecture Series on Information to collectively ponder short and long term goals;
- study dynamic information theory that extends information theory to time–space–varying situations;
- advance information algorithmics that develop new algorithms and data structures for the application of information;
- encourage and facilitate interdisciplinary collaborations;
- provide scholarships and fellowships for the best students, and support the development of new interdisciplinary courses.