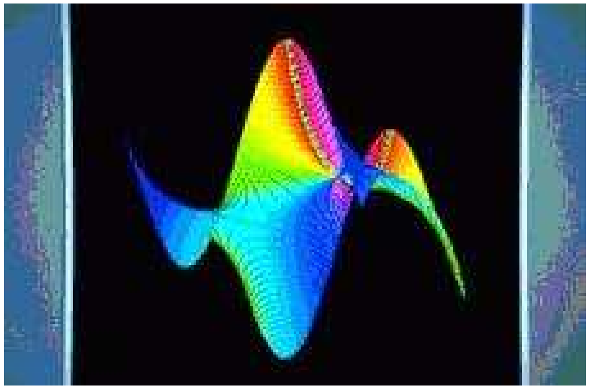
Analytic Information Theory and Beyond*

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

May 12, 2010

AofA and **IT** logos



Frankfurt 2010

*Research supported by NSF Science & Technology Center, and Humboldt Foundation.
†Joint work with M. Drmota, P. Flajolet, and P. Jacquet.

Outline

PART I: Shannon Information Theory

1. Shannon Legacy

2. Analytic Information Theory

3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources (Sequences mod 1)
   (b) Universal Memoryless Sources (Tree-like gen. func.)
   (c) Universal Markov Sources (Balance matrices)
   (d) Universal Renewal Sources (Combinatorial calculus)

PART II: Science of Information

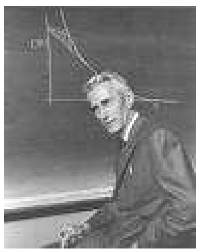
1. Post-Shannon Information
2. NSF Science and Technology Center

Shannon Legacy

The Information Revolution started in 1948, with the publication of:

A Mathematical Theory of Communication.

The digital age began.



Claude Shannon:
Shannon information quantifies the extent to which a recipient of data can reduce its statistical uncertainty.
"These semantic aspects of communication are irrelevant . . ."

Applications Enabler/Driver:

CD, iPod, DVD, video games, computer communication, Internet, Facebook, Google, . . .

Design Driver:

universal data compression, voiceband modems, CDMA, multiantenna, discrete denosing, space-time codes, cryptography, . . .

Three Theorems of Shannon

Theorem 1 & 3. (**Shannon 1948; Lossless & Lossy Data Compression**)

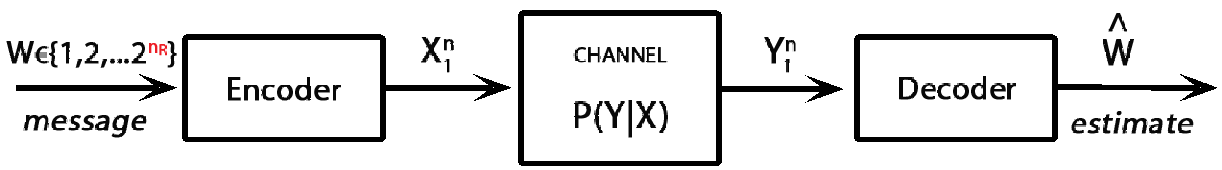
Lossless Compression: compression bit rate $\geq$ source entropy $H(X)$;

Lossy Compression: For distortion level D:
lossy bit rate $\geq$ rate distortion function $R(D)$

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.



Analytic Information Theory

- In the **1997 Shannon Lecture** Jacob Ziv presented compelling arguments for "backing off" from first-order asymptotics in order to predict the behavior of real systems with finite length description.

- To **overcome** these difficulties we propose replacing first-order analyses by full asymptotic expansions and more accurate analyses (e.g., large deviations, central limit laws).

- Following **Hadamard's precept**[1], we study information theory problems using techniques of complex analysis such as generating functions, combinatorial calculus, Rice's formula, Mellin transform, Fourier series, sequences distributed modulo 1, saddle point methods, analytic poissonization and depoissonization, and singularity analysis.

- This program, which applies complex-analytic tools to information theory, constitutes **analytic information theory**.

[1] The shortest path between two truths on the real line passes through the complex plane.

Outline Update

1. Shannon Legacy
2. Analytic Information Theory
3. Source Coding: The Redundancy Rate Problem

Source Coding

A source code is a bijective mapping

$$C : \mathcal{A}^* \to \{0,1\}^*$$

from sequences over the alphabet $\mathcal{A}$ to set $\{0,1\}^*$ of binary sequences.

The basic problem of source coding (i.e., data compression) is to find codes with shortest descriptions (lengths) either on *average* or for *individual sequences*.

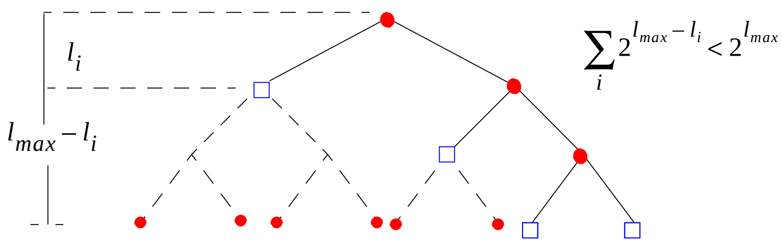
For a probabilistic source model $\mathcal{S}$ and a code C_n we let:

- $P(x_1^n)$ be the probability of $x_1^n = x_1 \ldots x_n$;
- $L(C_n, x_1^n)$ be the code length for x_1^n;
- Entropy $H_n(P) = -\sum_{x_1^n} P(x_1^n) \lg P(x_1^n)$.
- Fractional part: $\langle x \rangle = x - \lfloor x \rfloor$.

Prefix Codes

Prefix code is such that no codeword is a prefix of another codeword.

Kraft's Inequality
A code is a prefix code iff codeword lengths $\ell_1, \ell_2, \ldots, \ell_N$ satisfy the inequality



$$\sum_{i=1}^{N} 2^{-\ell_i} \leq 1.$$

Barron's lemma: For any sequence a_n of positive constants satisfying $\sum_n 2^{-a_n} < \infty$

$$\Pr\{L(X) < -\log P(X) - a_n\} \leq 2^{-a_n},$$

and therefore

$$L(X) \geq -\log P(X) - a_n \quad (\text{a.s}).$$

Shannon First Theorem: For any prefix code the average code length $\mathbf{E}[L(C_n, X_1^n)]$ cannot be smaller than the entropy of the source $H_n(P)$, that is,

$$\mathbf{E}[L(C_n, X_1^n)] \geq H_n(P).$$

Outline Update

1. Shannon Legacy
2. Analytic Information Theory
3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources

Redundancy

Known Source P: The pointwise redundancy $R_n(C_n, P; x_1^n)$ and the average redundancy $\bar{R}_n(C_n, P)$ are defined as

$$\begin{aligned} R_n(C_n, P; x_1^n) &= L(C_n, x_1^n) + \lg P(x_1^n) \\ \bar{R}_n(C_n) &= \mathbf{E}[L(C_n, X_1^n)] - H_n(P) \geq 0 \end{aligned}$$

The maximal or worst case redundancy is

$$R^*(C_n, P) = \max_{x_1^n}\{R_n(C_n, P; x_1^n)\}(\geq 0).$$

Huffman Code:

$$\bar{R}_n(P) = \min_{C_n \in \mathcal{C}} \mathbf{E}_{x_1^n}[L(C_n, x_1^n) + \log_2 P(x_1^n)].$$

Generalized Shannon Code: Drmota and W.S. (2001) consider

$$R_n^*(P) = \min_{C_n} \max_{x_1^n}[L(C_n, x_1^n) + \lg P(x_1^n)]$$

which is solved by for some constant s_0:

$$L(C_n^{GS}, x_1^n) = \begin{cases} \lfloor \lg 1/P(x_1^n) \rfloor & \text{if } \langle -\lg P(x_1^n) \rangle \leq s_0 \\ \lceil \lg 1/P(x_1^n) \rceil & \text{if } \langle -\lg P(x_1^n) \rangle > s_0 \end{cases}$$

Main Result

Theorem 1 (W.S., 2000). *Consider the Huffman block code of length n over a binary memoryless source with $p < \frac{1}{2}$. Then as $n \to \infty$*

$$\bar{R}_n^H = \begin{cases} \frac{3}{2} - \frac{1}{\ln 2} + o(1) \approx 0.057304 & \alpha \quad \text{irrational}, \\ \frac{3}{2} - \frac{1}{M}\left(\langle \beta M n \rangle - \frac{1}{2}\right) - \frac{1}{M(1-2^{-1/M})} 2^{-\langle n\beta M \rangle / M} + O(\rho^n) & \alpha = \frac{N}{M} \end{cases}$$

where $\gcd(N, M) = 1$, $\alpha = \log(1-p)/p$, $\beta = -\log(1-p)$, *and* $\rho < 1$.

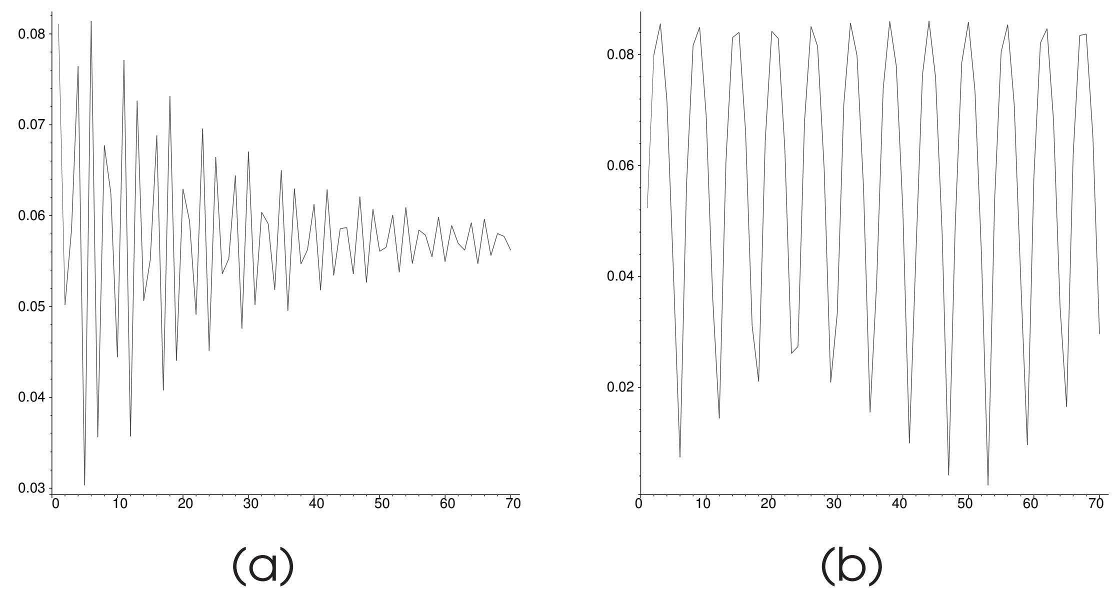


Figure 1: The average redundancy of Huffman codes versus block size n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

Why Two Modes: Shannon Code

Consider the Shannon code that assigns the length

$$L(C_n^S, x_1^n) = \lceil -\lg P(x_1^n) \rceil$$

where $P(x_1^n) = p^k(1-p)^{n-k}$, with p being **known** probability of generating 0 and k is the number of 0s.

The Shannon code redundancy is

$$\begin{aligned}
\bar{R}_n^S &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left(\lceil -\log_2(p^k(1-p)^{n-k}) \rceil + \log_2(p^k(1-p)^{n-k}) \right) \\
&= 1 - \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \langle \alpha k + \beta n \rangle \\
&= \begin{cases} \frac{1}{2} + o(1) & \alpha = \log_2(1-p)/p \quad \textbf{irrational} \\ \frac{1}{2} - \frac{1}{M}\left(\langle Mn\beta \rangle - \frac{1}{2} \right) + O(\rho^n) & \alpha = \frac{N}{M} \quad \textbf{rational} \end{cases}
\end{aligned}$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x, and

$$\alpha = \log_2\left(\frac{1-p}{p}\right), \quad \beta = \log_2\left(\frac{1}{1-p}\right).$$

Sketch of Proof: Sequences Modulo 1

To analyze redundancy for known sources one needs to understand asymptotic behavior of the following sum

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle)$$

for fixed p and some Riemann integrable function $f : [0,1] \to \mathbf{R}$.

The proof follows from the following two lemmas.
Lemma 2. *Let $0 < p < 1$ be a fixed real number and α be an irrational number. Then for every Riemann integrable function $f : [0,1] \to \mathbb{R}$*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle) = \int_0^1 f(t)\, dt,$$

where the convergence is uniform for all shifts $y \in \mathbb{R}$.
Lemma 3. *Let $\alpha = \frac{N}{M}$ be a rational number with $\gcd(N, M) = 1$. Then for bounded function $f : [0,1] \to \mathbf{R}$*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle) = \frac{1}{M} \sum_{l=0}^{M-1} f\left(\frac{l}{M} + \frac{\langle My \rangle}{M} \right) + O(\rho^n)$$

uniformly for all $y \in \mathbb{R}$ and some $\rho < 1$.

Outline Update

1. Shannon Legacy

2. Analytic Information Theory

3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources
   (b) Minimax Redundancy
   (c) Universal Memoryless Sources
   (d) Universal Markov Sources
   (e) Universal Renewal Sources

Minimax Redundancy

Unknown Source P

In practice, one can only hope to have some knowledge about a family of sources $\mathcal{S}$ that generates real data.

Following Davisson we define the average minimax redundancy $\bar{R}_n(\mathcal{S})$ and the worst case (maximal) minimax redundancy $R_n^*(\mathcal{S})$ for a family of sources $\mathcal{S}$ as

$$\begin{aligned}
\bar{R}_n(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \mathbf{E}[L(C_n, x_1^n) + \lg P(x_1^n)] \\
R_n^*(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)].
\end{aligned}$$

In the minimax scenario we look for the best code for the the worst source.

Source Coding Goal:
Find data compression algorithms that **match optimal redundancy rates** either on average or for individual sequences.

Maximal Minimax Redundancy

We consider the following **classes of sources** $\mathcal{S}$:

- **Memoryless sources** $\mathcal{M}_0$ over an m-ary alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$, that is,
$$P(x_1^n) = {p_1}^{k_1} \cdots {p_m}^{k_m}$$
with $k_1 + \cdots + k_m = n$, where p_i are unknown!

- **Markov sources** $\mathcal{M}_r$ over an m-ary alphabet of order r
$$P(x_1^n) = p_{11}^{k_{11}} \cdots p_{ij}^{k_{ij}} \cdots p_{mm}^{k_{mm}}$$
where k_{ij} is the number of pair symbols ij in x_1^n, and they satisfy the **balance property**. Notice that p_{ij} are unknown.

- **Renewal Sources** $\mathcal{R}_0$ where an 1 is introduced after a run of 0s distributed according to some distribution.

(Improved) Shtarkov Bounds for R_n^*

For the maximal minimax redundancy define

$$Q^*(x_1^n) := \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{\sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)}.$$

the maximum likelihood distribution. Observe that

$$\begin{aligned}
R_n^*(\mathcal{S}) &= \min_{C_n\in\mathcal{C}} \sup_{P\in\mathcal{S}} \max_{x_1^n} (L(C_n, x_1^n) + \lg P(x_1^n)) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \sup_{P\in\mathcal{S}} \lg P(x_1^n) \right) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg \sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)] \\
&= R_n^{GS}(Q^*) + \lg \sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)
\end{aligned}$$

where $R_n^{GS}(Q^*)$ is the maximal redundancy of a generalized Shannon code built for the (known) distribution Q^*. We also write

$$D_n(\mathcal{S}) = \lg \left(\sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(x_1^n) \right) := \lg d_n(\mathcal{S}).$$

Outline Update

1. Shannon Legacy

2. Analytic Information Theory

3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources

Maximal Minimax for Memoryless Sources

We first consider the maximal minimax redundancy $R_n^*(\mathcal{M}_0)$ for a class of memoryless sources over a finite m-ary alphabet. Observe that

$$\begin{aligned} d_n(\mathcal{M}_0) &= \sum_{x_1^n} \sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} \\ &= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} \\ &= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \left(\frac{k_1}{n}\right)^{k_1}\cdots\left(\frac{k_m}{n}\right)^{k_m}. \end{aligned}$$

The **summation set** is

$$I(k_1,\ldots,k_m) = \{(k_1,\ldots,k_m) : \ k_1+\cdots+k_m = n\}.$$

The number $N_{\mathbf{k}}$ of **types** $\mathbf{k} = (k_1,\ldots,k_m)$ is

$$N_{\mathbf{k}} = \binom{n}{k_1,\ldots,k_m}$$

The (unnormalized) **likelihood distribution** is

$$\sup_{p_1,\ldots,p_m} p_1^{k_1}\cdots p_m^{k_m} = \left(\frac{k_1}{n}\right)^{k_1}\cdots\left(\frac{k_m}{n}\right)^{k_m}$$

Generating Function for $d_n(\mathcal{M}_0)$

We write

$$d_n(\mathcal{M}_0) = \frac{n!}{n^n} \sum_{k_1+\cdots+k_m=n} \frac{k_1^{k_1}}{k_1!} \cdots \frac{k_m^{k_m}}{k_m!}$$

Let us introduce a tree-generating function

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k = \frac{1}{1 - T(z)},$$

where $T(z)$ satisfies $T(z) = ze^{T(z)}$ $(= -W(-z)$, **Lambert's** W-function) and also

$$T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} z^k$$

enumerates all rooted labeled trees. Let now

$$D_m(z) = \sum_{n=0}^{\infty} z^n \frac{n^n}{n!} d_n(\mathcal{M}_0).$$

Then by the convolution formula

$$D_m(z) = [B(z)]^m.$$

Asymptotics

The function $B(z)$ has an algebraic singularity at $z = e^{-1}$ (it becomes a multi-valued function) and one finds

$$B(z) = \frac{1}{\sqrt{2(1-ez)}} + \frac{1}{3} + O(\sqrt{(1-ez)}.$$

The singularity analysis yields (cf. Clarke & Barron, 1990, W.S., 1998)

$$\begin{aligned} D_n(\mathcal{M}_0) \quad &= \quad \frac{m-1}{2}\log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2}-\frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{n}} \\ &+ \quad \left(\frac{3+m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2}-\frac{1}{2})}\right) \cdot \frac{1}{n} + \cdots \end{aligned}$$

To complete the analysis, we need $\bar{R}_n^{GS}(Q^*)$. Drmota & W.S., 2001 proved

$$R_n^{GS}(Q^*) = -\frac{\ln\frac{1}{m-1}\ln m}{\ln m} + o(1),$$

In general, the term $o(1)$ can not be improved. Thus

$$R_n^*(\mathcal{M}_0) = \frac{m-1}{2}\log\left(\frac{n}{2}\right) - \frac{\ln\frac{1}{m-1}\ln m}{\ln m} + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + o(1).$$

Outline Update

1. Shannon Legacy
2. Analytic Information Theory
3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources

Maximal Minimax for Markov Sources

(i) $\mathcal{M}_1$ is a Markov source of order $r = 1$,
(ii) the transition matrix $P = \{p_{ij}\}_{i,j=1}^{m}$
(iii) easy to see that

$$d_n(\mathcal{M}_1) \quad = \quad \sum_{x_1^n} \sup_P \; p_{11}^{k_{11}} \cdots p_{mm}^{k_{mm}} = \sum_{\mathbf{k} \in \mathcal{F}_n} \mathbf{M_k} \left(\frac{k_{11}}{k_1}\right)^{k_{11}} \cdots \left(\frac{k_{mm}}{k_m}\right)^{k_{mm}},$$

k_{ij} is the number of pairs $ij \in \mathcal{A}^2$ in x_1^n, $k_i = \sum_{j=1}^m k_{ij}$ such that

$$\mathcal{F}_n : \sum_{i,j=1}^{m} k_{ij} = n, \quad \text{and} \quad \sum_{j=1}^{m} k_{ij} = \sum_{j=0}^{m-1} k_{ji},$$

Matrix $\mathbf{k}$ satisfying the above conditions is called the frequency matrix or **Markov type**. $M_{\mathbf{k}}$ represents the numbers of strings x_1^n of **type** $\mathbf{k}$.

For circular strings (i.e., after the n symbol we re-visit the first symbol of x_1^n), the frequency matrix $[k_{ij}]$ satisfies the following constraints that we denote as $\mathcal{F}_n$

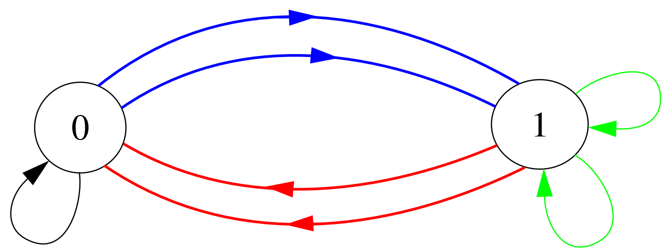
$$\sum_{1 \le i,j \le m} k_{ij} = n, \qquad \sum_{j=1}^{m} k_{ij} = \sum_{j=1}^{m} k_{ji}, \;\forall\, i \;\; (\textbf{balance property})$$

Markov Types and Eulerian Cycles

Let $\mathbf{k} = [k_{ij}]_{i,j=1}^{m}$ be a Markov type satisfying $\mathcal{F}_n$ (balance property).

Example: Let $\mathcal{A} = \{0, 1\}$ and

$$\mathbf{k} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$



Two questions:

A: *How many sequences of a given type* $\mathbf{k}$ *are there?*
How many Eulerian paths in the underlying multigraph over $|\mathcal{A}|$ *with* k_{ij} *edges are there?*

B: How many distinct matrices $\mathbf{k}$ satisfying $\mathcal{F}_n$ are there?
How many Markov types $\mathcal{P}_n(m)$ are there?
How many integer solutions to the balance equations are there?

Main Technical Tool

Let $g_{\mathbf{k}}$ be a sequence of scalars indexed by matrices $\mathbf{k}$ and

$$g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$$

be its regular generating function, and

$$\mathcal{F}g(\mathbf{z}) = \sum_{\mathbf{k}\in\mathcal{F}} g_{\mathbf{k}} z^{\mathbf{k}} = \sum_{n\geq 0} \sum_{\mathbf{k}\in\mathcal{F}_n} g_{\mathbf{k}} z^{\mathbf{k}}$$

the $\mathcal{F}$-generating function of $g_{\mathbf{k}}$ for which $\mathbf{k} \in \mathcal{F}$.
Lemma 5. *Let* $g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$. *Then*

$$\mathcal{F}g(\mathbf{z}) := \sum_{n\geq 0} \sum_{\mathbf{k}\in\mathcal{F}_n} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \left(\frac{1}{2\jmath\pi}\right)^m \oint \frac{dx_1}{x_1} \cdots \oint \frac{dx_m}{x_m} g([z_{ij}\frac{x_j}{x_i}])$$

with the ij*-th coefficient of* $[z_{ij}\frac{x_j}{x_i}]$ *is* $z_{ij}\frac{x_j}{x_i}$.

Proof. It suffices to observe

$$g([z_{ij}\frac{x_j}{x_i}]) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} \prod_{i=1}^{m} x_i^{\sum_i k_{ij} - \sum_j k_{ij}}$$

Thus $\mathcal{F}g(\mathbf{z})$ is the coefficient of $g([z_{ij}\frac{x_j}{x_i}])$ at $x_1^0 x_2^0 \cdots x_m^0$.

Number of Markov Types and Sequences of a Given Type

(i) The number of strings $N_{\mathbf{k}}^{a,b}$ of type $\mathbf{k}$ that start with an a and ends with a b is

$$N_{\mathbf{k}}^{b,a} = \frac{k_{ba}}{k_b} B_{\mathbf{k}} \cdot \det_{bb}(\mathbf{I} - \mathbf{k}^*)$$

where $\mathbf{k}^*$ is the normalized matrix such that $\mathbf{k}^* = [k_{ij}/k_i]$ and

$$B_{\mathbf{k}} = \binom{k_1}{k_{11} \cdots k_{1m}} \cdots \binom{k_m}{k_{m1} \cdots k_{mm}}.$$

(ii) For fixed m and $n \to \infty$ the number of Markov types is

$$|\mathcal{P}_n(m)| = d(m) \frac{n^{m^2-m}}{(m^2-m)!} + O(n^{m^2-m-1})$$

where $d(m)$ is a constant that also can be expressed as

$$d(m) = \frac{1}{(2\pi)^{m-1}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{(m-1)-fold} \prod_{j=1}^{m-1} \frac{1}{1+\varphi_j^2} \cdot \prod_{k \neq \ell} \frac{1}{1+(\varphi_k - \varphi_\ell)^2} d\varphi_1 d\varphi_2 \cdots d\varphi_{m-1}.$$

(iii) For $m \to \infty$, Provided $m^4 = o(n)$, we find

$$|\mathcal{P}_n(m)| \sim \frac{\sqrt{2} m^{3m/2} e^{m^2}}{m^{2m^2} 2^m \pi^{m/2}} \cdot n^{m^2-m}.$$

Markov Redundancy: Main Results

Theorem 3 (Jacquet and W.S., 2004). *Let $\mathcal{M}_1$ be a Markov source over an m-ry alphabet. Then*

$$d_n(\mathcal{M}_1) = \left(\frac{n}{2\pi}\right)^{m(m-1)/2} A_m \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

with

$$A_m = \int_{\mathcal{K}(1)} m F_m(y_{ij}) \prod_i \frac{\sqrt{\sum_j y_{ij}}}{\prod_j \sqrt{y_{ij}}} d[y_{ij}]$$

where $\mathcal{K}(1) = \{y_{ij} : \sum_{ij} y_{ij} = 1\}$ and $F_m(\cdot)$ is a polynomial of degree $m - 1$.

In particular, for $m = 2$ $A_2 = 16 \times$ Catalan where Catalan is Catalan's constant $\sum_i \frac{(-1)^i}{(2i+1)^2} \approx 0.915965594$.

Theorem 4. *Let $\mathcal{M}_r$ be a Markov source of order r. Then*

$$d_n(\mathcal{M}_r) = \left(\frac{n}{2\pi}\right)^{m^r(m-1)/2} A_m^r \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

where A_m^r is a constant defined in a similar fashion as A_m above.

Outline Update

1. Shannon Legacy
2. Analytic Information Theory
3. Source Coding: The Redundancy Rate Problem
   (a) Known Sources
   (b) Universal Memoryless Sources
   (c) Universal Markov Sources
   (d) Universal Renewal Sources

Renewal Sources

The **renewal process** defined as follows:

- Let $T_1, T_2 \ldots$ be a sequence of i.i.d. positive-valued random variables with distribution $Q(j) = \Pr\{T_i = j\}$.
- The process $T_0, T_0 + T_1, T_0 + T_1 + T_2, \ldots$ is called the renewal process.
- In a **binary renewal sequence** the positions of the $1'$s are at the renewal epochs (runs of zeros) $T_0, T_0 + T_1, \ldots$.
- We start with $x_0 = 1$.

Csiszár and Shields (1996) proved that $R_n^*(\mathcal{R}_0) = \Theta(\sqrt{n})$.

We prove the following result.
Theorem 5 (Flajolet and W.S., 1998). *Consider the class of renewal processes. Then*

$$R_n^*(\mathcal{R}_0) = \frac{2}{\log 2}\sqrt{cn} + O(\log n).$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$.

Maximal Minimax Redundancy

For a sequence

$$x_0^n = 10^{\alpha_1}10^{\alpha_2}1\cdots 10^{\alpha_n}1\underbrace{0\cdots 0}_{k^*}$$

k_m is the number of i such that $\alpha_i = m$. Then

$$P(x_1^n) = Q^{k_0}(0)Q^{k_1}(1)\cdots Q^{k_{n-1}}(n-1)\Pr\{T_1 > k^*\}.$$

It can be proved that

$$r_{n+1} - 1 \le d_n(\mathcal{R}_0) \le \sum_{m=0}^{n} r_m$$

where

$$\begin{aligned} r_n &= \sum_{k=0}^{n} r_{n,k} \\ r_{n,k} &= \sum_{\mathcal{P}(n,k)} \binom{k}{k_0 \ldots k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \cdots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}} \end{aligned}$$

where $\mathcal{P}(n,k)$ is is the partition of n into k terms, i.e.,

$$\begin{aligned} n &= k_0 + 2k_1 + \cdots + nk_{n-1}, \\ k &= k_0 + \cdots + k_{n-1}. \end{aligned}$$

Main Results

Theorem 6 (Flajolet and W.S., 1998). *We have the following asymptotics*

$$\log r_n = \frac{2}{\log 2}\sqrt{cn} - \frac{5}{8}\lg n + \frac{1}{2}\lg\log n + O(1)$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$*.*

Asymptotic analysis is sophisticated and follows these steps:

- first, we transform r_n into another quantity s_n.
- use combinatorial calculus to find the generating function of s_n (infinite product of tree-functions $B(z)$);
- transform this product into a harmonic sum that can be analyzed asymptotically by the Mellin transform;
- asymptotic expansion of the generating function around $z = 1$.
- finally, estimate $R_n^*(\mathcal{R}_0)$ by the saddle point method.

Asymptotics: The Main Idea

The quantity r_n is to hard to analyze due to the factor $k!/k^k$, hence we define a new quantity s_n defined as

$$\begin{cases} s_n & = & \sum_{k=0}^{n} s_{n,k} \\ s_{n,k} & = & e^{-k} \sum_{\mathcal{P}(n,k)} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!}. \end{cases}$$

To analyze it, we introduce the random variable K_n as follows

$$\Pr\{K_n = k\} = \frac{s_{n,k}}{s_n}.$$

Stirling's formula yields

$$\begin{aligned} \frac{r_n}{s_n} &= \sum_{k=0}^{n} \frac{r_{n,k}}{s_{n,k}} \frac{s_{n,k}}{s_n} = \mathbf{E}[(K_n)! K_n^{-K_n} e^{-K_n}] \\ &= \mathbf{E}[\sqrt{2\pi K_n}] + O(\mathbf{E}[K_n^{-\frac{1}{2}}]). \end{aligned}$$

Fundamental Lemmas

Lemma 6. *Let* $\mu_n = \mathbf{E}[K_n]$ *and* $\sigma_n^2 = \mathbf{Var}(K_n)$.

$$\begin{aligned} s_n &\sim \exp\left(2\sqrt{cn} - \frac{7}{8}\log n + d + o(1)\right) \\ \mu_n &= \frac{1}{4}\sqrt{\frac{n}{c}}\log\frac{n}{c} + o(\sqrt{n}) \\ \sigma_n^2 &= O(n\log n) = o(\mu_n^2), \end{aligned}$$

where $c = \pi^2/6 - 1$, $d = -\log 2 - \frac{3}{8}\log c - \frac{3}{4}\log\pi$.

Lemma 7. *For large* n

$$\begin{aligned} \mathbf{E}[\sqrt{K_n}] &= \mu_n^{1/2}(1 + o(1)) \\ \mathbf{E}[K_n{}^{-\frac{1}{2}}] &= o(1). \end{aligned}$$

where $\mu_n = \mathbf{E}[K_n]$.

Thus

$$\begin{aligned} r_n &= s_n\mathbf{E}[\sqrt{2\pi K_n}](1 + o(1)) \\ &= s_n\sqrt{2\pi\mu_n}(1 + o(1)). \end{aligned}$$

Sketch of a Proof: Generating Functions

1. Define the function $\beta(z)$ as

$$\beta(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} e^{-k} z^k.$$

One has (e.g., by Lagrange inversion or otherwise)

$$\beta(z) = \frac{1}{1 - T(ze^{-1})}.$$

2. Define

$$S_n(u) = \sum_{k=0}^{\infty} s_{n,k} u^k, \qquad S(z,u) = \sum_{n=0}^{\infty} S_n(u) z^n.$$

Since $s_{n,k}$ involves convolutions of sequences of the form $k^k/k!$, we have

$$\begin{aligned} S(z,u) &= \sum_{\mathcal{P}_{n,k}} z^{1k_0+2k_1+\cdots} \left(\frac{u}{e}\right)^{k_0+\cdots+k_{n-1}} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!} \\ &= \prod_{i=1}^{\infty} \beta(z^i u). \end{aligned}$$

We need to compute $s_n = [z^n]S(z,1)$, coefficient at z^n of $S(z,1)$.

Mellin Asymptotics

3. Let $L(z) = \log S(z,1)$ and $z = e^{-t}$, so that

$$L(e^{-t}) = \sum_{k=1}^{\infty} \log \beta(e^{-kt}).$$

Mellin transform techniques.

4. The Mellin transform

$$L^*(s) = \int_0^\infty L(e^{-t}) x^{s-1} dx$$

by the harmonic sum property: $\mathcal{M}\left(\sum_{k\geq 0} \lambda_k g(\mu_k x)\right) = g^*(s) \sum_{k\geq 0} \lambda_k \mu_k^{-s}$:

$$L^*(s) = \zeta(s)\Lambda(s), \quad \Re(s) \in (1, \infty)$$

where $\zeta(s) = \sum_{n\geq 1} n^{-s}$ is the Riemann zeta function, and

$$\Lambda(s) = \int_0^\infty \log \beta(e^{-t}) t^{s-1} dt.$$

This leads to

$$L^*(s) \asymp \left(\frac{\Lambda(1)}{s-1}\right)_{s=1} + \left(-\frac{1}{4s^2} - \frac{\log \pi}{4s}\right)_{s=0}.$$

What's Next?

5. An application of the converse mapping property (M4) allows us to come back to the original function,

$$L(e^{-t}) = \frac{\Lambda(1)}{t} + \frac{1}{4}\log t - \frac{1}{4}\log \pi + O(\sqrt{t}),$$

which translates in

$$L(z) = \frac{\Lambda(1)}{1-z} + \frac{1}{4}\log(1-z) - \frac{1}{4}\log \pi - \frac{1}{2}\Lambda(1) + O(\sqrt{1-z}).$$

where

$$\begin{aligned} c = \Lambda(1) &= -\int_0^1 \log(1 - T(x/e))\frac{dx}{x} \\ &= \frac{\pi^2}{6} - 1. \end{aligned}$$

6. In summary, we just proved that, as $z \to 1^-$,

$$S(z,1) = e^{L(z)} = a(1-z)^{\frac{1}{4}}\exp\left(\frac{c}{1-z}\right)(1+o(1)),$$

where $a = \exp(-\frac{1}{4}\log \pi - \frac{1}{2}c)$.

7. To extract asymptotic we need to apply the **saddle point method**.

Outline Update

1. Shannon Information Theory
2. Source Coding
3. The Redundancy Rate Problem

PART II: Science of Information

1. What is Information?
2. Post-Shannon Information
3. NSF Science and Technology Center

Post-Shannon Challenges

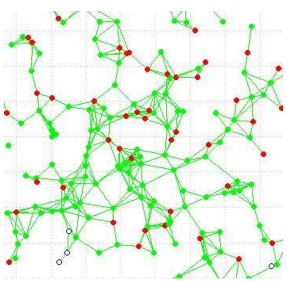
Classical Information Theory needs a recharge to meet new challenges of nowadays applications in biology, modern communication, knowledge extraction, economics and physics, ...

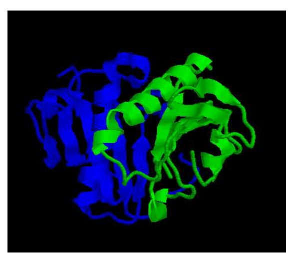
We need to extend Shannon information theory to include new aspects of information such as:

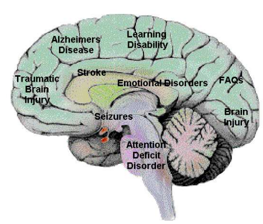
structure, time, space, and semantics ,

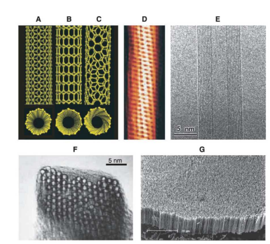
and others such as:

dynamic information, limited resources, complexity, physical information, representation-invariant information, and cooperation & dependency.



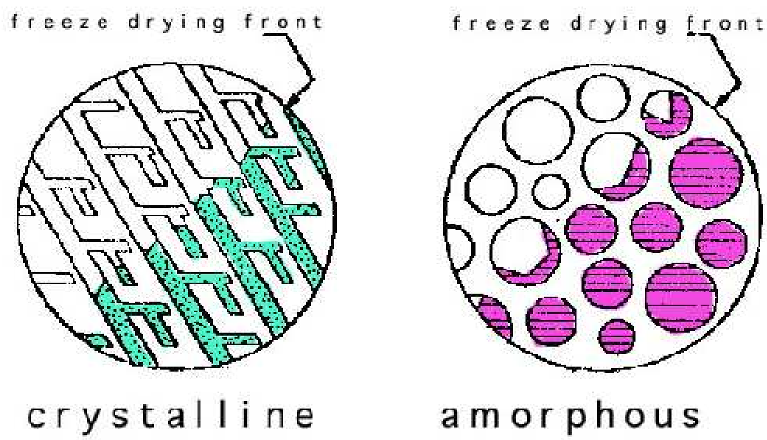




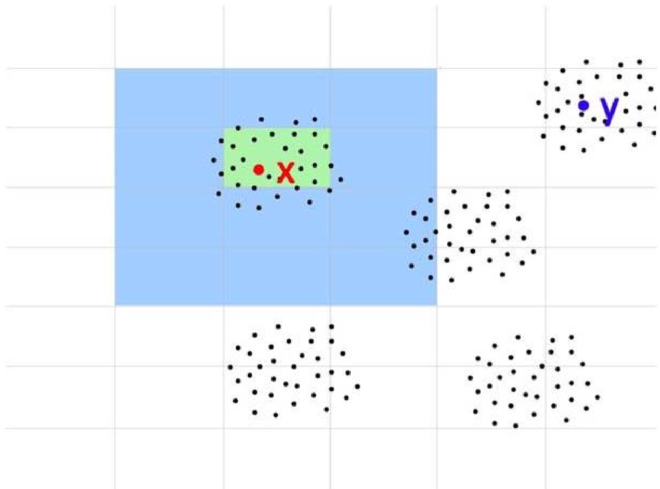


Structure, Time & Space, and Semantics

Structure:
Measures are needed for quantifying information embodied in structures
(e.g., material structures, nanostructures, biomolecules, gene regulatory networks protein interaction networks, social networks, financial transactions).



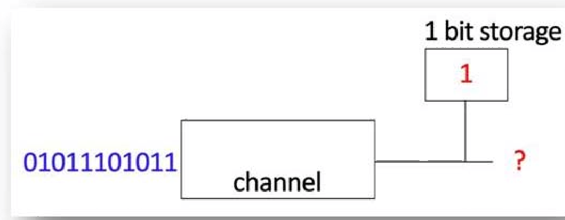
Time & Space:
Classical Information Theory is at its weakest in dealing with problems of delay
(e.g., information arriving late maybe useless or has less value).



Semantics & Learnable information:
Data driven science focuses on extracting information from data. How much information can actually be extracted from a given data repository? How much knowledge is in Google's database?

Limited Resources, Representation, and Cooperation

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



Representation-invariant of information: How to know whether two representations of the same information are information equivalent?

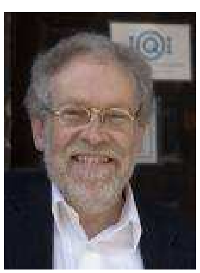
Cooperation. Often subsystems may be in conflict (e.g., denial of service) or in collusion (e.g., price fixing). How does cooperation impact information? (In wireless networks nodes should cooperate in their own self-interest.)

Standing on the Shoulders of Giants . . .

F. Brooks, jr. (JACM, 2003, "Three Great Challenges . . ."):
We have **no theory** 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.

Manfred Eigen (Nobel Prize, 1967)
"The differentiable characteristic of the living systems is Information.
Information assures the controlled reproduction of all constituents,
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**".

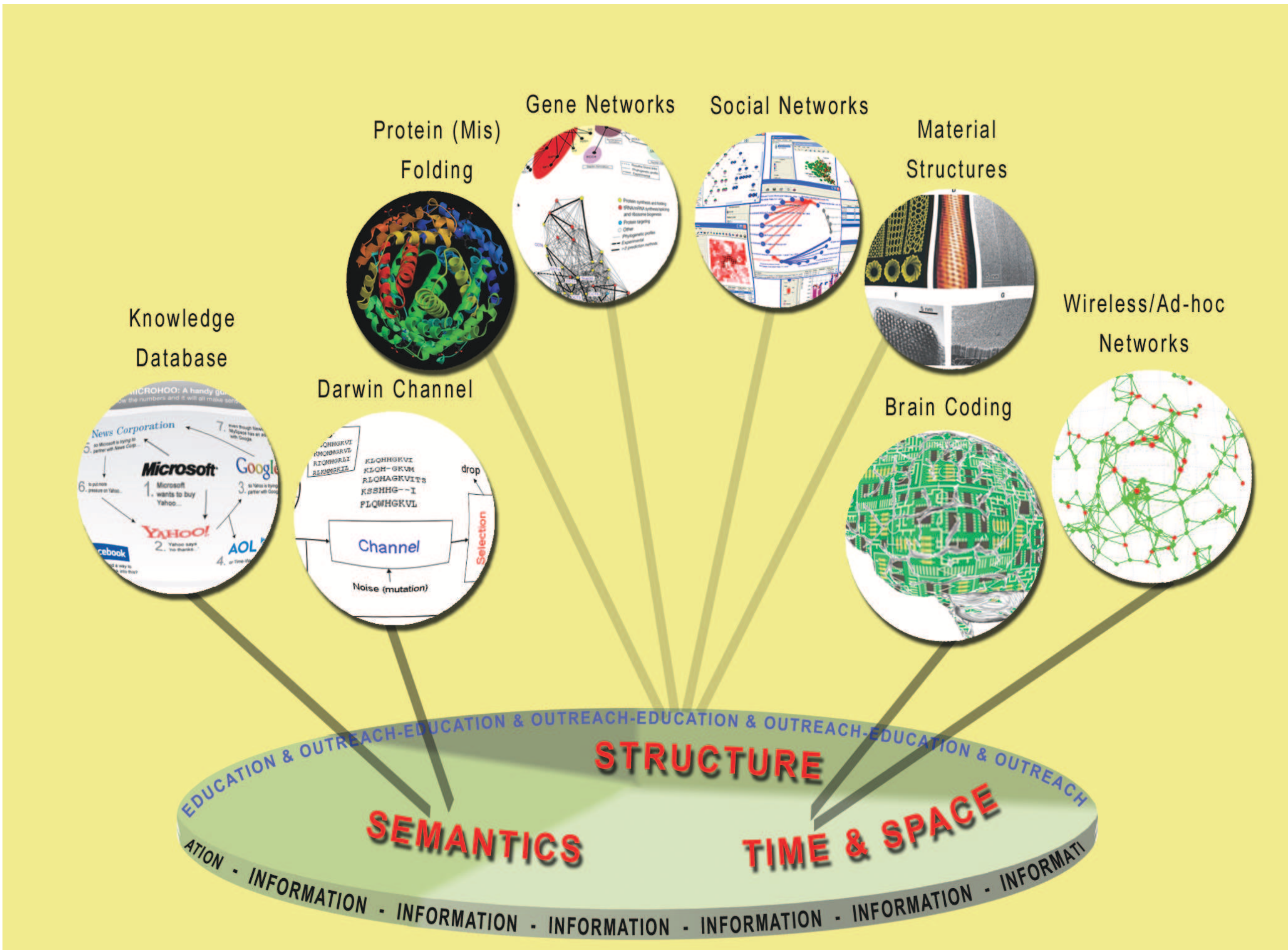
P. Nurse, (Nature, 2008, "Life, Logic, and Information"):
Focusing on information flow will help to understand better cell
how cells and organisms work.
". . . the generation of spatial and temporal order,
memory and reproduction are not fully understood".



A. Zeilinger (Nature, 2005)
. . . reality and information are two sides of the same coin,
that is, they are in a deep sense indistinguishable.

Science of Information

The overarching vision of **Science of Information** is to develop rigorous principles guiding the extraction, manipulation, and exchange of information, integrating elements of space, time, structure, and semantics.



Institute for Science of Information

In 2008 at Purdue we launched the

Institute for Science of Information

and in 2010 National Science Foundation established $25M

Science and Technology Center

at Purdue to do ccollaborative work with Berkeley, MIT, Princeton, Stanford, UIUC and Bryn Mawr & Howard U. integrating research and teaching activities aimed at investigating the role of **information** from various viewpoints: from the fundamental theoretical underpinnings of greeninformation 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:

- develope post-Shannon Information Theory,
- Prestige Science Lecture Series on Information to collectively ponder short and long term goals;
- organize meetings and workshops (e.g., Information Beyond Shannon, Orlando 2005, and Venice 2008).
- initiate similar world-wide centers supporting research on information.

THANK YOU