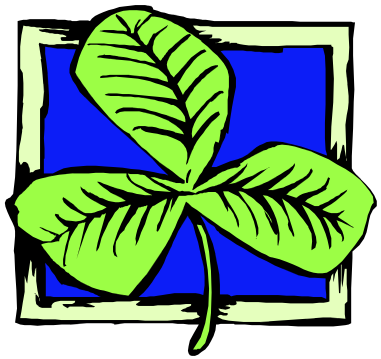
DIMACS Wkshp on Network Info. Theory
St. Patrick's Day 2003

Network Information Theory — Some Tentative Definitions



James L. Massey

Prof.-em. ETH Zurich
Adjunct Prof., Lund Univ.
Trondhjemsgade 3, 2TH
DK-2100 Copenhagen East

JamesMassey@compuserve.com

Reminder of Gallager's notation:

$\mathbf{X}^N = X_1, X_2, \ldots, X_N$ (sequence of N random variables)

$\mathbf{x}^N = x_1, x_2, \ldots, x_N$ (a realization of $\mathbf{X}^N$)

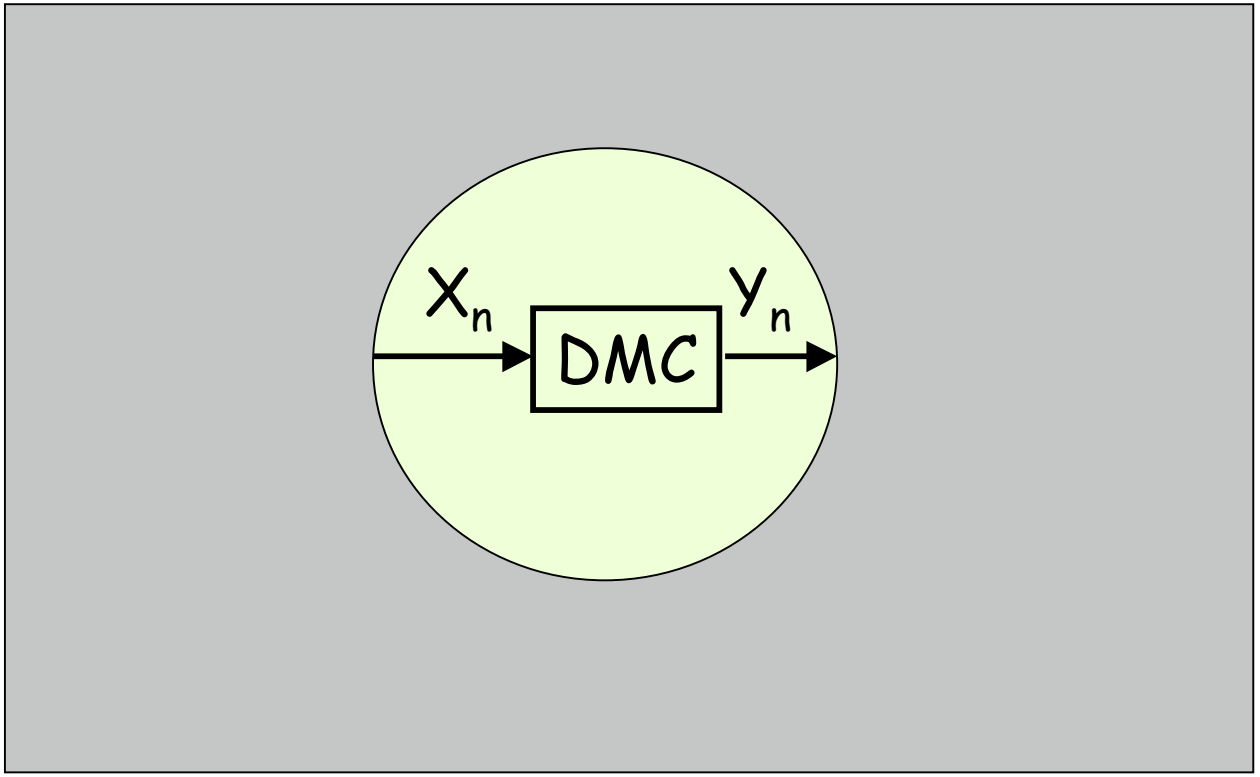
$P(\mathbf{x}^N) = P_{X^N}(\mathbf{x}^N) = \Pr(\mathbf{X}^N = \mathbf{x}^N)$

$P(\mathbf{y}^N | \mathbf{x}^N) = P_{Y^N|X^N}(\mathbf{y}^N | \mathbf{x}^N) = \Pr(\mathbf{Y}^N = \mathbf{y}^N | \mathbf{X}^N = \mathbf{x}^N)$

Some additional notation:

$*$ = "concatenation of sequences"

$\Longrightarrow 0 * \mathbf{y}^{N-1} = \mathbf{0}, Y_1, Y_2, \ldots, Y_{N-1}$



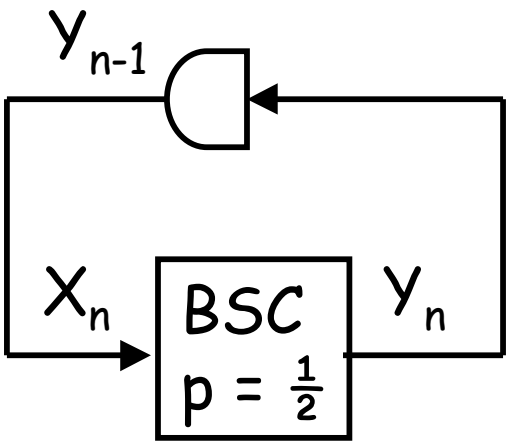
What is the probability law of a Discrete Memoryless Channel (DMC)?

$$P(\mathbf{y}^N \mid \mathbf{x}^N) = \prod_{n=1}^{N} P_{Y|X}(y_n|x_n), \text{ all } n \geq 1 \quad ???$$

Does it matter what is in the gray area?

The **natural and least restrictive assumption** to make about the gray area is that it contains **no negative delays** and **no closed paths through boxes for which the path delay is zero**.

This is the kind of assumption made in the study of discrete-time systems.



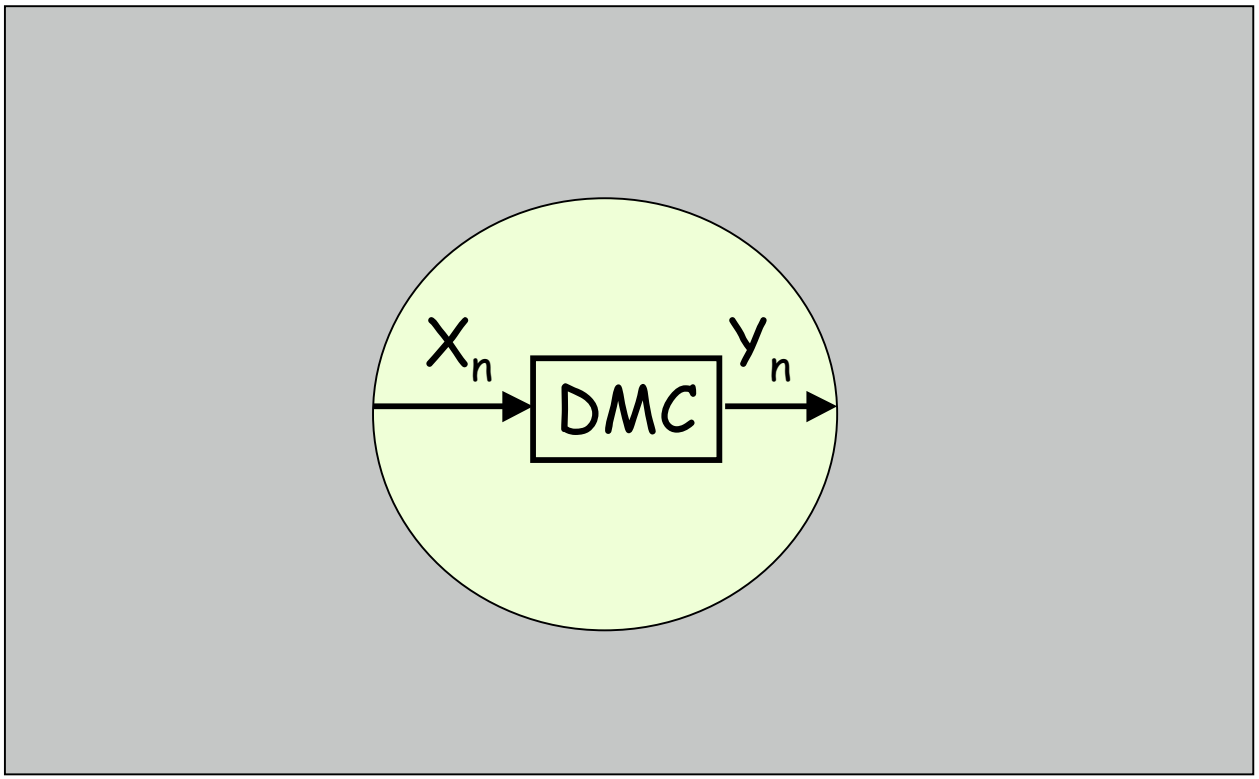
Let $Y_0 = 0$ be the initial condition in the delay.

N.B.: $X_n = Y_{n-1}$, all $n \geq 1$

$\Longrightarrow \quad P(Y_1Y_2 = 1\ 0 |\ X_1X_2 = 0\ 0) = 0$

But $P_{Y|X}(1|0)P_{Y|X}(0|0) = 1/4$

What is wrong with our definition of a DMC?



Here is the "correct" probability law of a DMC !

$$P(y_n | \mathbf{x}^n \mathbf{y}^{n-1}) = P_{Y|X}(y_n|x_n), \text{ all } n \geq 1$$

This means that X_n is a **sufficient statistic** for reasoning about Y_n given the **observation** $(\mathbf{X}^n, \mathbf{Y}^{n-1})$.

In his book [R. Ash, *Information Theory*. New York: Wiley Interscience, 1965], Bob Ash correctly gave the probability law for the DMC but "spoiled" his definition by further requiring the **"causality condition"** that, for $1 \le n \le N$,

$$P(y_n | x^N y^{n-1}) = P(y_n | x^n y^{n-1}).$$

Adding this condition gives

$$P(y^N | x^N) = \prod_{n=1}^{N} P(y_n | x^N y^{n-1}) = \prod_{n=1}^{N} P(y_n | x^n y^{n-1})$$

$$= \prod_{n=1}^{N} P(y_n | x_n).$$

Ash's **"causality condition"** has nothing to do with causality! What it does do is to **prohibit feedback**!

Note that Ash's "causality condition"

$P(y_n | \mathbf{x}^N \mathbf{y}^{n-1}) = P(y_n | \mathbf{x}^n \mathbf{y}^{n-1})$ for $1 \leq n \leq N$

can equivalently be written as

$H(Y_n | \mathbf{X}^N \mathbf{y}^{n-1}) = H(Y_n | \mathbf{X}^n \mathbf{y}^{n-1})$ for $1 \leq n \leq N$.

It is often convenient to work with uncertainties (discrete entropies) rather than with the probability distributions themselves.

A more natural definition for prohibiting feedback is to say that **a channel is used without feedback** if

$P(x_n | \mathbf{x}^{n-1} \mathbf{y}^{n-1}) = P(x_n | \mathbf{x}^{n-1})$ for all $n \geq 1$.

To show that these two conditions for prohibiting the use of feedback are equivalent, it suffices to shows that both give the same $P(x^N y^N)$.

Using the "natural condition", we have

$$P(x_n y_n \mid \mathbf{x}^{n-1} \mathbf{y}^{n-1}) = P(x_n \mid \mathbf{x}^{n-1} \mathbf{y}^{n-1})\, P(y_n \mid \mathbf{x}^n \mathbf{y}^{n-1})$$

$$= P(x_n \mid \mathbf{x}^{n-1})\, P(y_n \mid \mathbf{x}^n \mathbf{y}^{n-1})$$

$$\Longrightarrow P(\mathbf{x}^N \mathbf{y}^N) = P(\mathbf{x}^N) \prod_{n=1}^{N} P(y_n \mid \mathbf{x}^n \mathbf{y}^{n-1}).$$

Using Ash's condition, we have

$$P(\mathbf{x}^N \mathbf{y}^N) = P(\mathbf{x}^N)\, P(\mathbf{y}^N \mid \mathbf{x}^N) = P(\mathbf{x}^N) \prod_{n=1}^{N} P(y_n \mid \mathbf{x}^N \mathbf{y}^{n-1})$$

$$= P(\mathbf{x}^N) \prod_{n=1}^{N} P(y_n \mid \mathbf{x}^n \mathbf{y}^{n-1}).$$

Unlike causality, probabilistic dependence has no direction. Whether *A* causes *B* or *B* causes *A*, *A* and *B* will be statistically dependent.

Essentially this is why $I(X^N ; Y^N) = I(Y^N ; X^N)$

or, equivalently,

$H(Y^N) - H(Y^N \mid X^N) = H(X^N) - H(X^N \mid Y^N)$.

My 1990 definition of **directed information**:

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} I(X^n ; Y_n | Y^{n-1}).$$

(By ignoring information that Y_n may be giving about future X digits, we are considering only the information flowing from X digits to Y digits.)

In terms of uncertainties (discrete entropies)

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(Y_n \mid Y^{n-1}) - H(Y_n \mid X^n Y^{n-1})].$$

We recall that

$$I(X^N ; Y^N) = \sum_{n=1}^{N} [H(Y_n \mid Y^{n-1}) - H(Y_n \mid X^N Y^{n-1}).$$

Equivalently, we can write

$$I(X^N \to Y^N) = \sum_{n=1}^{N} [H(X^n \mid Y^{n-1}) - H(X^n \mid Y^n)].$$

Because $H(Y_n \mid X^N Y^{n-1}) \leq H(Y_n \mid X^n Y^{n-1})$, it follows that

$$I(X^N \to Y^N) \leq I(X^N ; Y^N)$$

with equality if and only if there is no feedback, i.e., if and only if (natural condition) $P(x_n \mid x^{n-1} y^{n-1}) = P(x_n \mid x^{n-1})$ [or, equivalently, (Ash's condition) $P(y_n \mid x^N y^{n-1}) = P(y_n \mid x^n y^{n-1})$] for $1 \leq n \leq N$.

If $\mathbf{X}^N$ and $\mathbf{Y}^N$ are the input and output sequences of a DMC, then

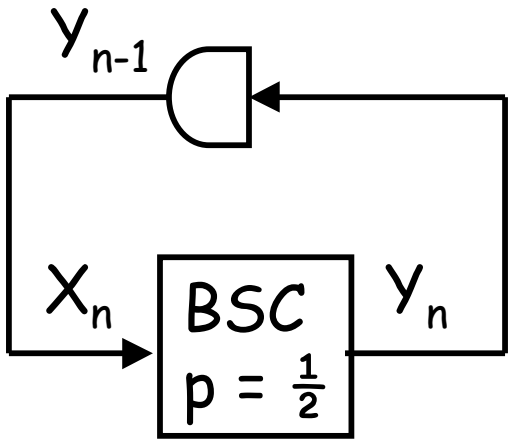
$$I(\mathbf{X}^N \rightarrow \mathbf{Y}^N) = \sum_{n=1}^{N} [H(Y_n \mid \mathbf{Y}^{n-1}) - H(Y_n \mid \mathbf{X}^n \mathbf{Y}^{n-1})]$$

$$= \sum_{n=1}^{N} [H(Y_n \mid \mathbf{Y}^{n-1}) - H(Y_n \mid X_n)]$$

$$\leq \sum_{n=1}^{N} [H(Y_n) - H(Y_n \mid X_n)] =$$

$$= \sum_{n=1}^{N} I(X_n ; Y_n)$$

with equality if and only if $Y_1, Y_2, \ldots, Y_N$ are statistically independent.



Let $Y_0 = 0$ be the initial contents of the delay.

N.B.: $X_n = Y_{n-1}$, all $n \geq 1$
so that $X_1 = 0$

$H(Y^N) = N$ bits

$H(Y^N \mid X^N) = H(Y_N \mid X^N) = 1$ bit

$\Rightarrow I(X^N ; Y^N) = N - 1$ bits

$I(X_n ; Y_n) = 0$ bits, all n

$\Rightarrow I(X^N \rightarrow Y^N) = 0$ bits

$H(X_n \mid X^{n-1}) = 1$ bit, all $n \geq 2$

$H(X_1) = 0$ bits

$H(X_n \mid O*Y^{n-1} X^{n-1}) = 0$, $n > 1$

$\Rightarrow I(O*Y^{N-1} \rightarrow X^N) = N - 1$ bits

Note that $I(X^N \rightarrow Y^N) + I(O*Y^{N-1} \rightarrow X^N) = I(X^N ; Y^N)$.

Conservation Law for Directed Information

$$I(X^N \to Y^N) + I(0 * Y^{N-1} \to X^N) = I(X^N ; Y^N).$$

Proof by induction:

$I(X^1 \to Y^1) = I(X_1 ; Y_1)$ and $I(0 \to X^1) = 0$.

Note that

$$I(X^{n+1} \to Y^{n+1}) = I(X^n \to Y^n) + I(X^{n+1}; Y_{n+1} | Y^n)$$

and similarly that

$$I(0 * Y^n \to X^{n+1}) = I(0 * Y^{n-1} \to X^n) + I(0 * Y^n ; X_{n+1} | X^n)$$
$$= I(0 * Y^{n-1} \to X^n) + I(Y^n ; X_{n+1} | X^n).$$

Thus, by the induction hypothesis,

$$I(X^{n+1} \to Y^{n+1}) + I(0 * Y^n \to X^{n+1}) = I(X^n ; Y^n) + I(X^{n+1}; Y_{n+1} | Y^n)$$
$$+ I(Y^n ; X_{n+1} | X^n)$$
$$= I(X^{n+1}; Y^n) + I(X^{n+1}; Y_{n+1} | Y^n)$$
$$= I(X^{n+1}; Y^{n+1}).$$

Gerhard Kramer's definition of **causal conditioning**:

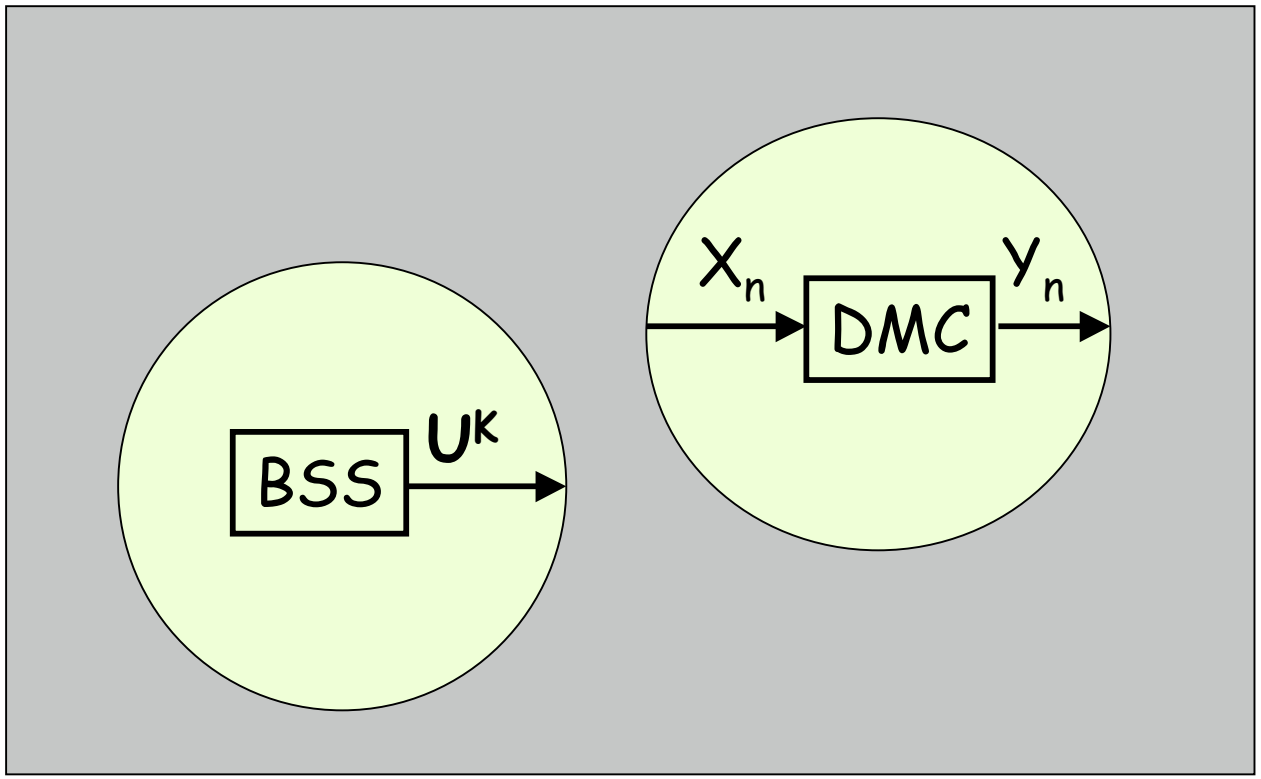
$$H(Y^N \| X^N) = \sum_{n=1}^{N} H(Y_n \mid X^n Y^{n-1})$$

In terms of causal conditioning, one can write

$$I(X^N \rightarrow Y^N) = H(Y^N) - H(Y^N \| X^N)$$

Gerhard also defined causally conditioned directed information in the following manner:

$$I(X^N \rightarrow Y^N \| Z^N) = H(Y^N \| Z^N) - H(Y^N \| X^N Z^N)$$



We want to consider a synchronized network in which all devices are governed by the same notion of time.

What we mean by **synchronization**:

$X_1, X_2, X_3, X_4, X_5, \ldots, X_N$
$Y_1, Y_2, Y_3, Y_4, Y_5, \ldots, Y_N$
$Z_1, Z_2, Z_3, Z_4, Z_5, \ldots, Z_N$

Three arbitrary **clocked sequences**

The meaning is that the n^{th} variable in each sequence, i.e., X_n, Y_n, and Z_n, take on their values at the same time instant. Moreover, the n^{th} variable takes on its value before the $(n+1)^{st}$ variable. Note also that Y_{n-1} is the n^{th} variable in the concatenated sequence $0 * \mathbf{Y}^{N-1}$.

Which kinds of sequences should be clocked?

Channels: Input and output sequences are clocked.

Delays: Input and output sequences are clocked.

Sources: **Not clocked** - output present at creation!

Channel Encoders: Only output sequence and input feedback sequence (if present) are clocked.

Channel Decoders: Only input sequence and output feedback sequence (if present) are clocked.

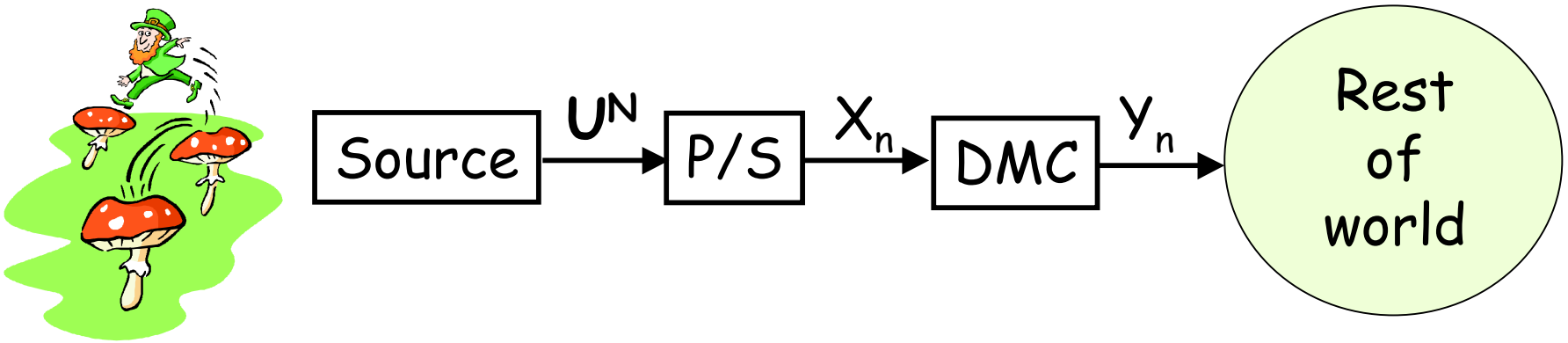
Source Encoders: **Not clocked** - input and output present at creation!

Source Reconstructors: **Not clocked** - input and output present at end of world!

Clocked inputs to network components can come only from clocked outputs of other network components.

$\Rightarrow$ you cannot connect a source directly to a channel! You must use a channel encoder.

Michael Gastpar's direct transmission of a source over a channel has to be understood as using a parallel-to-series converter as the channel encoder.



The **causality assumption** for synchronized networks:

For any channel (whether or not memoryless), the input-output sequences (X^n, Y^{n-1}) are a **sufficient statistic** for reasoning about the output Y_n given the **observation** (X^n, Y^{n-1}, W), where W is any random quantity composed of inputs and outputs of other network elements at time n or earlier, $1 \le n \le N$.

Thus, for any **source output** U^K,

$$H(Y_n | X^n Y^{n-1} U^K) = H(Y_n | X^n Y^{n-1}).$$

$$\Rightarrow H(Y^N | U^K) = \sum_{n=1}^{N} H(Y_n | U^K Y^{n-1}) \ge \sum_{n=1}^{N} H(Y_n | U^K X^n Y^{n-1}) =$$

$$\sum_{n=1}^{N} H(Y_n | X^n Y^{n-1}) = H(Y^N \| X^N)$$

Thus, $I(U^K ; Y^N) \le I(X^N \to Y^N)$.

An immediate consequence of the fact that

$$I(\mathbf{U}^K ; \mathbf{Y}^N) \leq I(\mathbf{X}^N \rightarrow \mathbf{Y}^N) \leq \sum_{n=1}^{N} I(X_n ; Y_n) \leq N\, C_{DMC}$$

is that **feedback does not increase the capacity of a discrete memoryless channel (DMC)**.

How can one logically prove this result if one uses the "usual definition" of a DMC, which is given on slide 2 and which in fact prohibits the use of feedback?

How can we reason correctly about complicated networks of sources, channel, encoders, decoders and delays without some sort of careful statement of our assumptions similar to that given here today?

Some references:

J. L. Massey, "Causality, Feedback and Directed Information," pp. 303-305 in *Proc. 1990 Int. Symp. on Info. Th. & its Appls.*, Hawaii, USA, Nov. 27-30, 1990.

H. Marko, "The Bidirectional Communication Theory – A Generalization of Information Theory", *IEEE Trans. Commun.*, vol. COM-21, pp. 1345-1351, Dec. 1973.

G. Kramer, *Directed Information for Channels with Feedback*, ETH Series in Inform. Proc., vol. 11. Konstanz: Hartung--Gorre, 1998.

G. Kramer, "Capacity Results for the Discrete Memoryless Network," *IEEE Trans. Inform. Th.*, vol. IT-49, pp. 4-21, Jan. 2003.