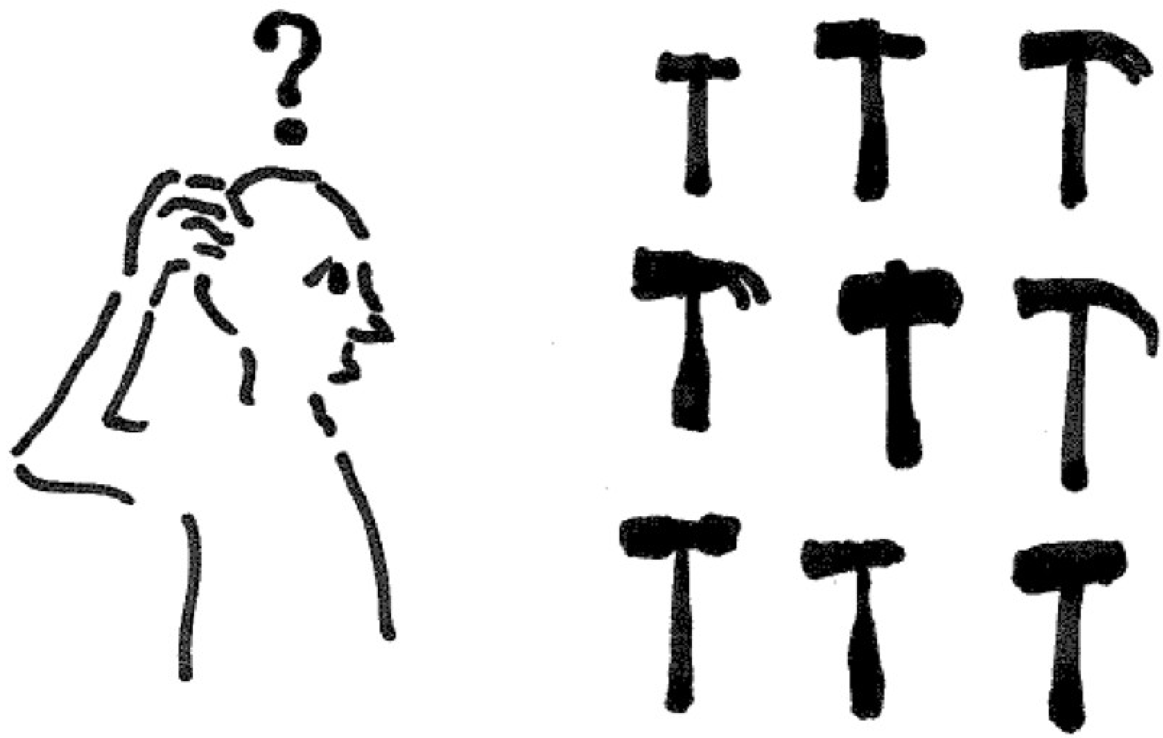
A Quantitative Measure of Relevance Based on Kelly Gambling Theory



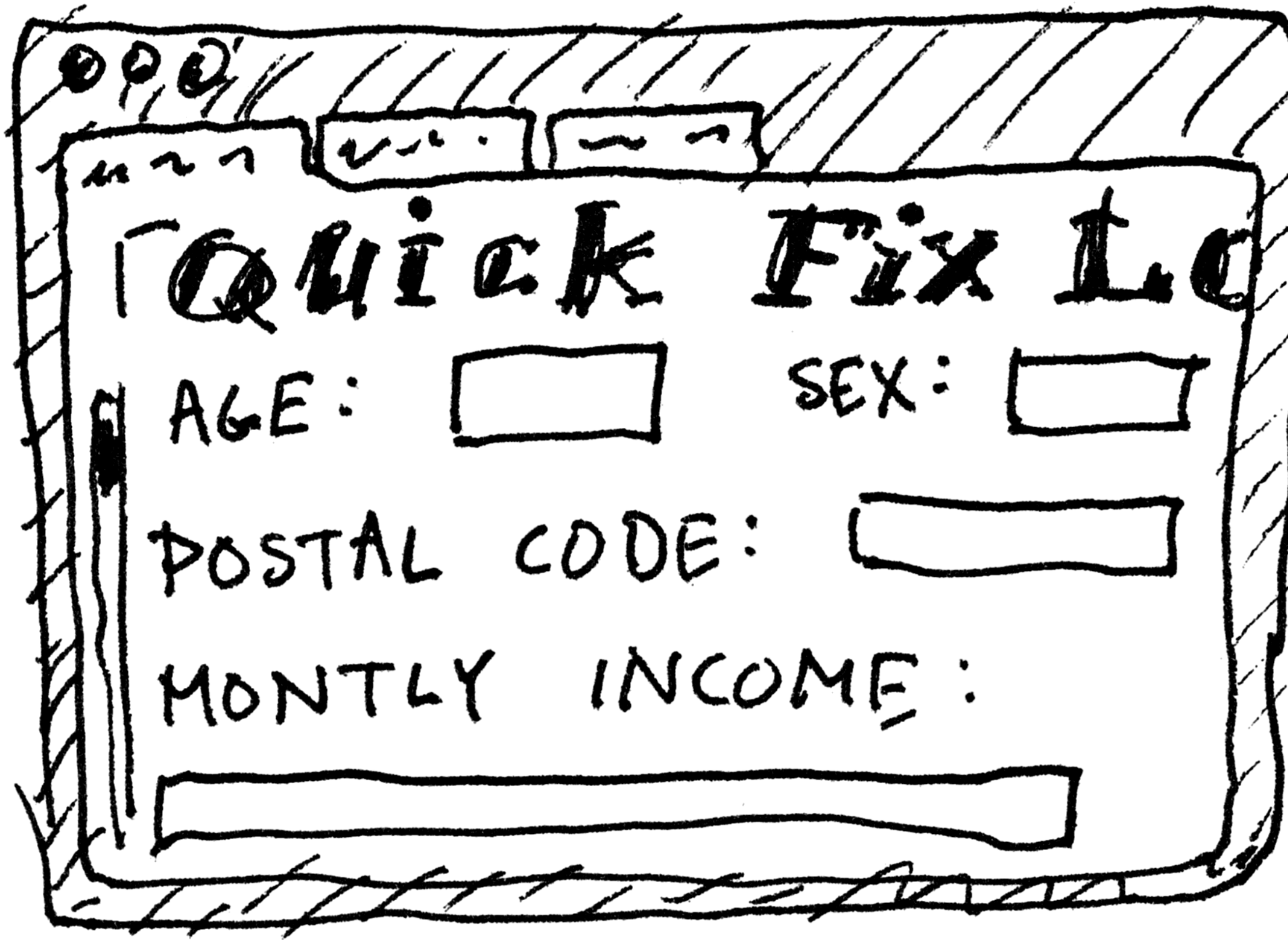
Mathias Winther Madsen

Institute for Logic, Language, and Computation
University of Amsterdam

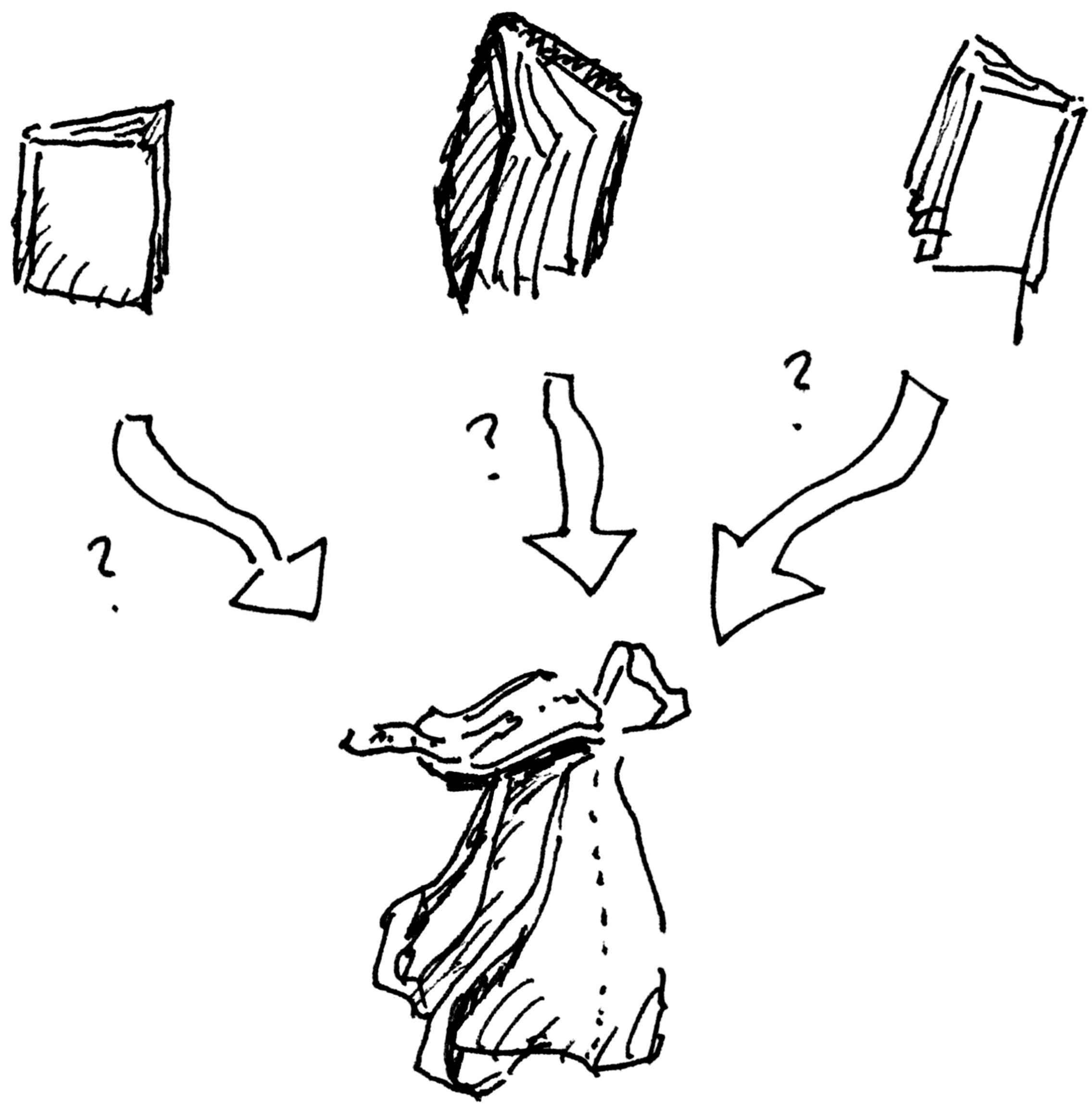
PLAN

- Why?
- How?
- Examples

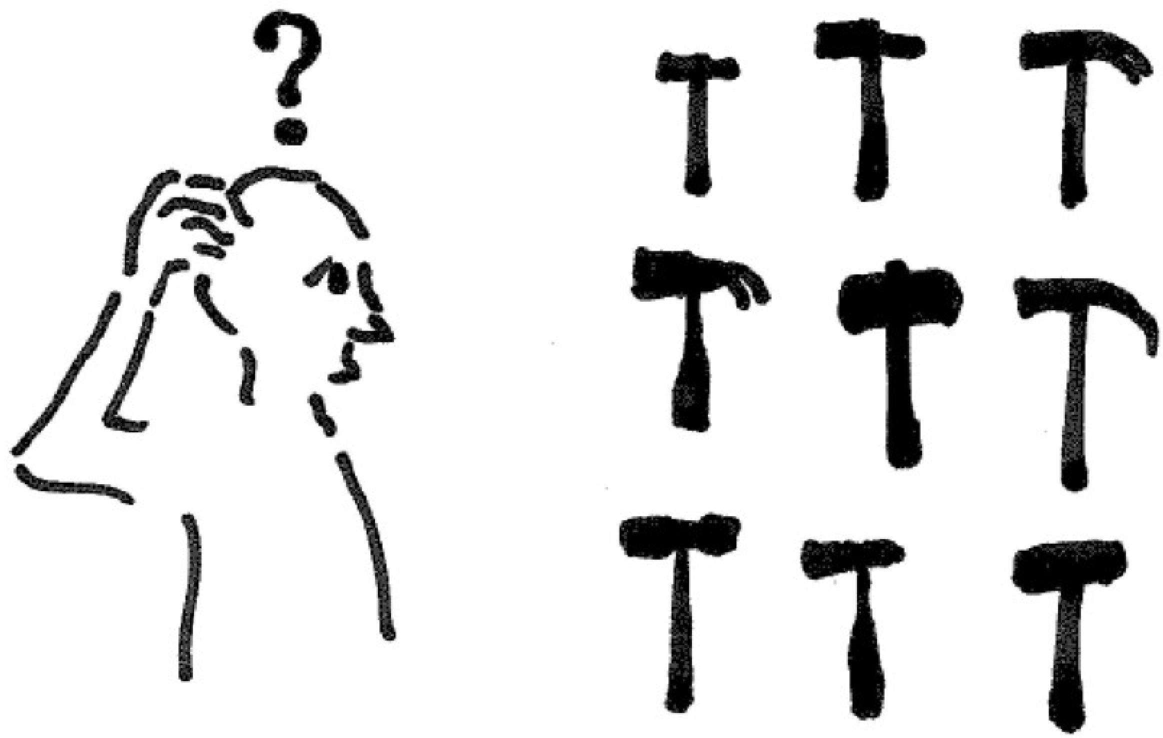
Why?



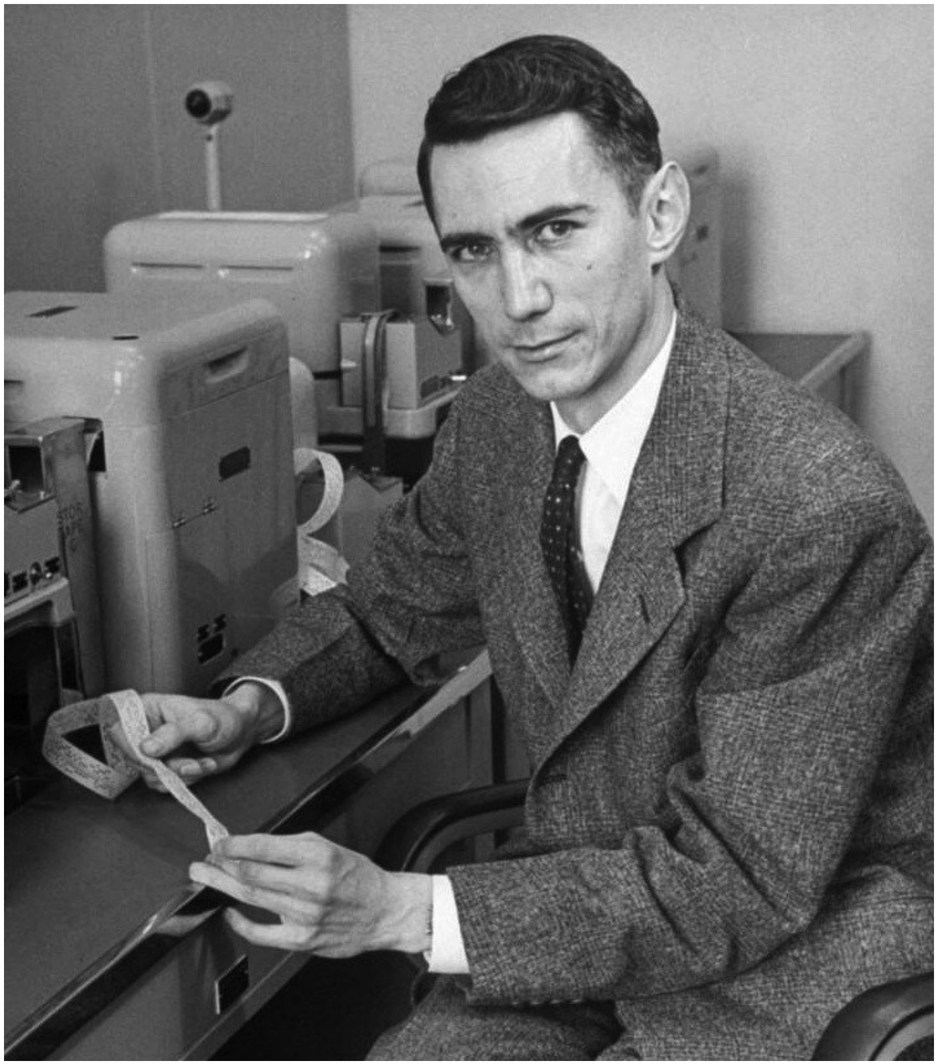
Why?



How?



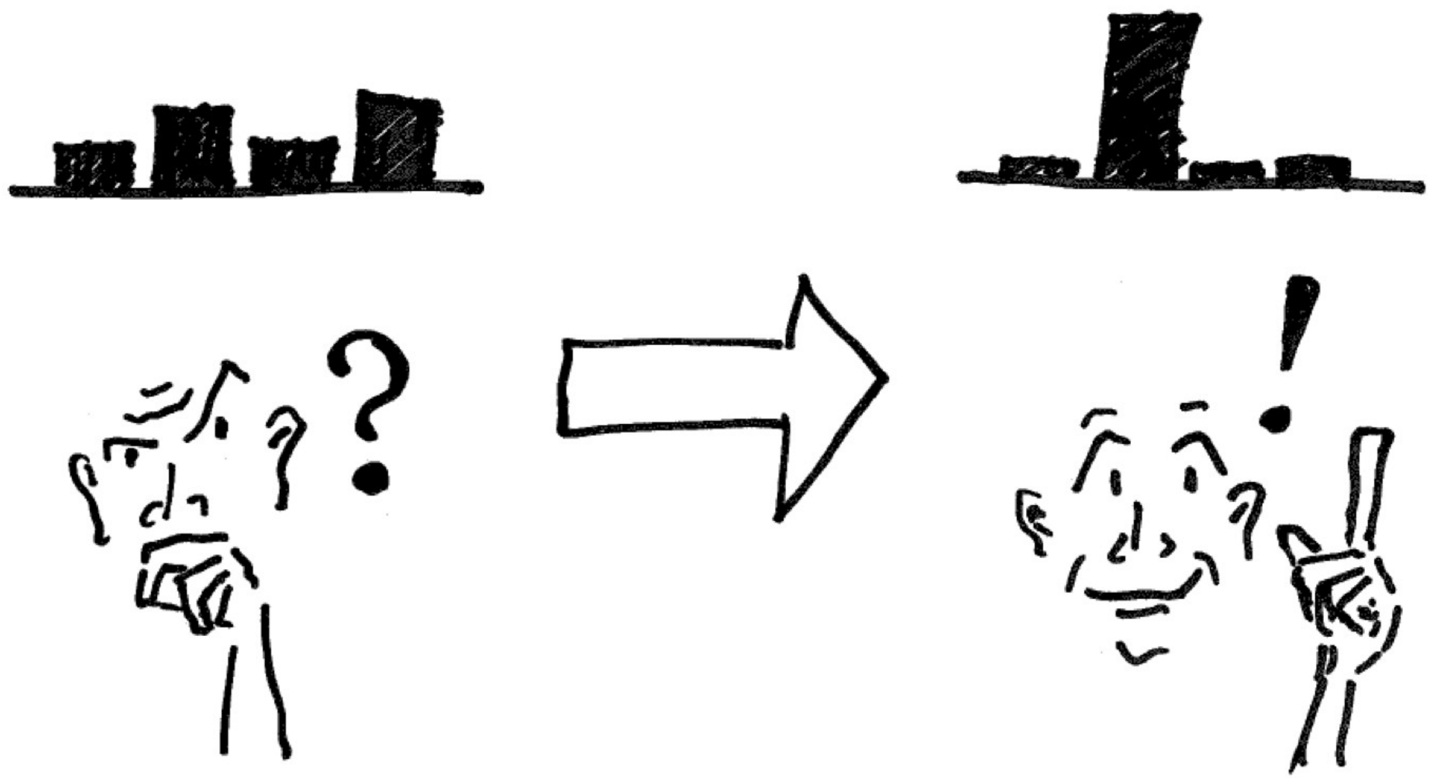
Why not use Shannon information?



Claude Shannon
(1916 – 2001)

$$H(X) = E\left[\log \frac{1}{\Pr(X = x)}\right]$$

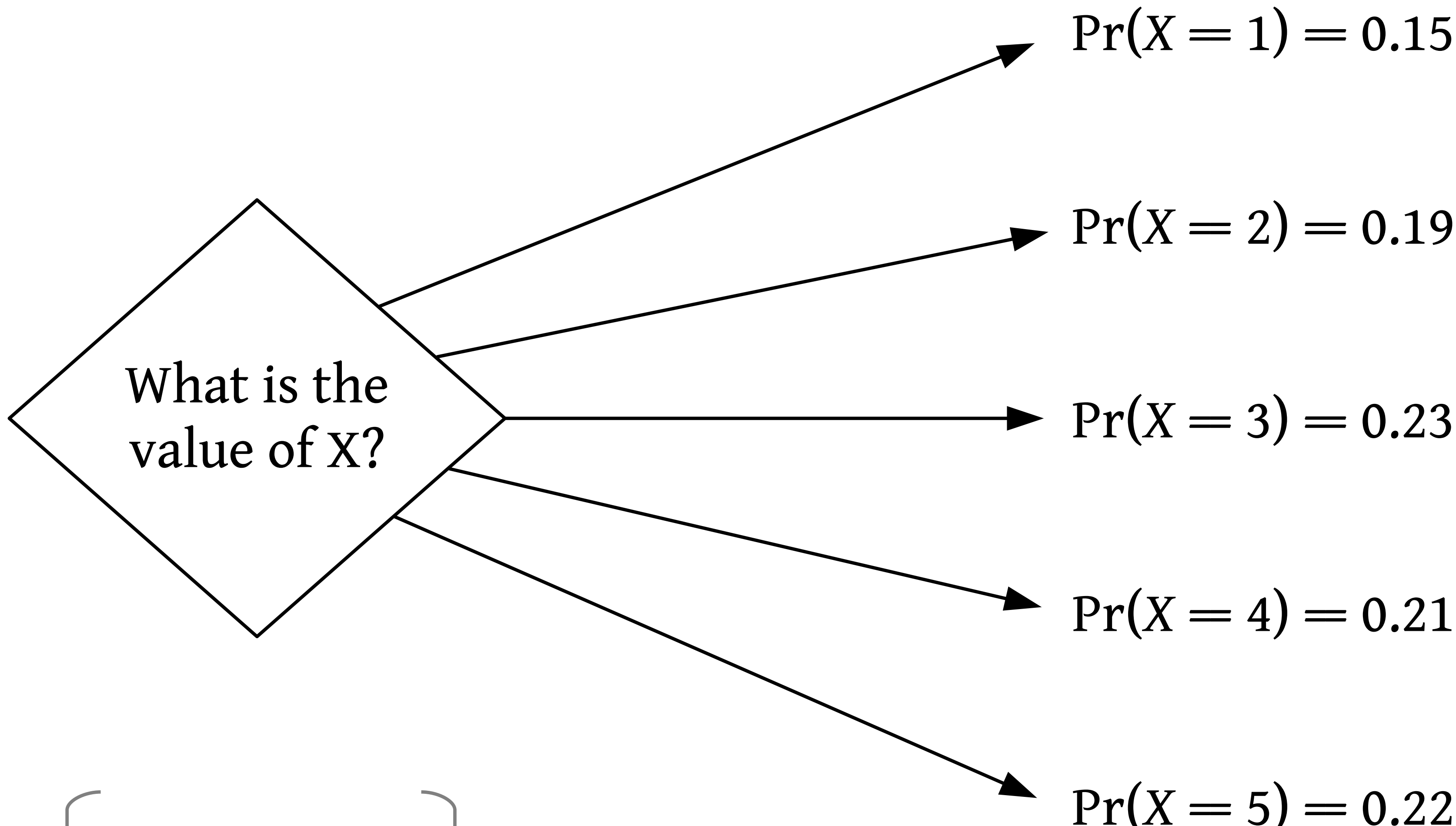
Why not use Shannon information?



Information Content = Prior Uncertainty − Posterior Uncertainty

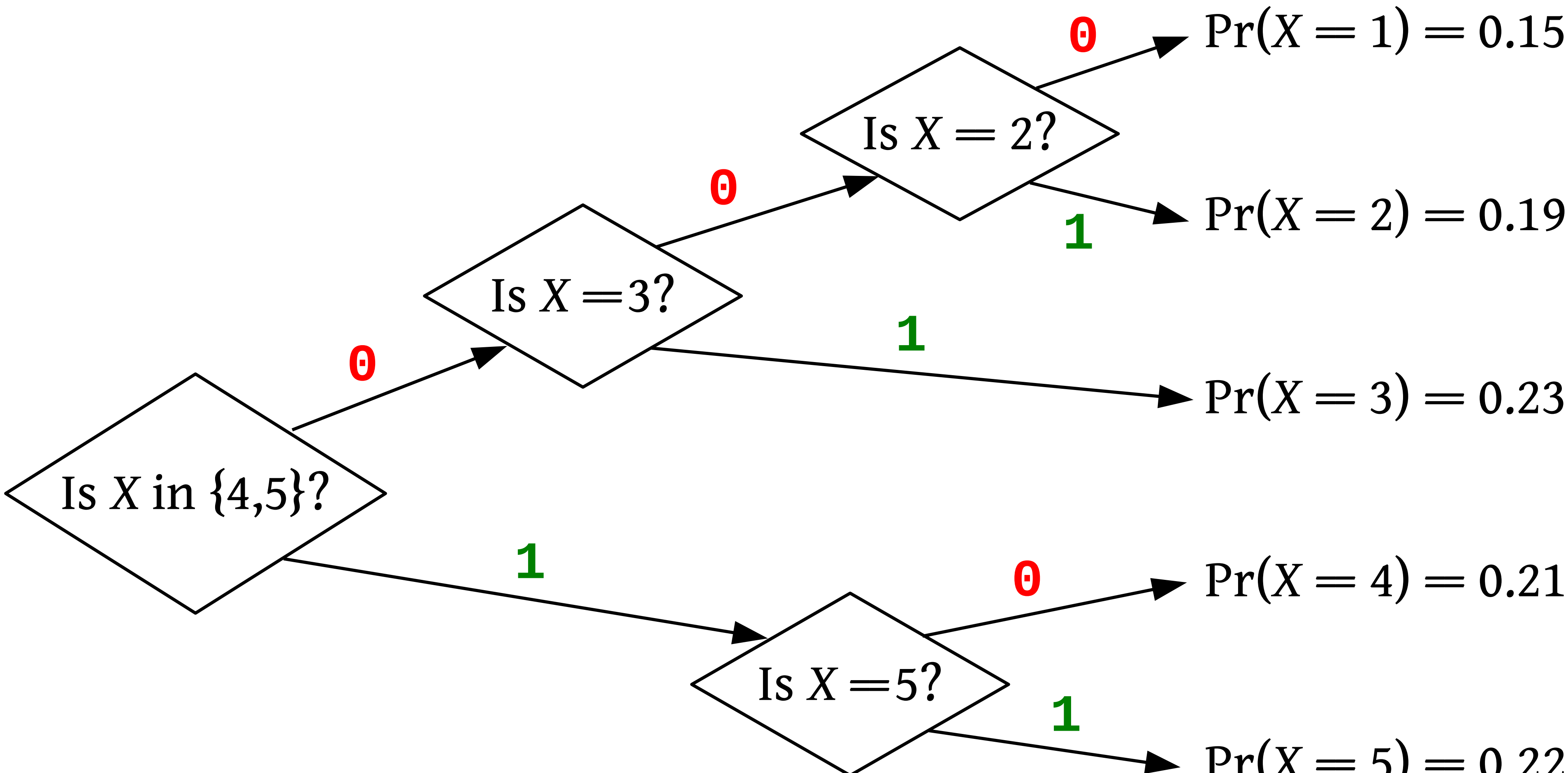
(cf. Klir 2008; Shannon 1948)

Why not use Shannon information?



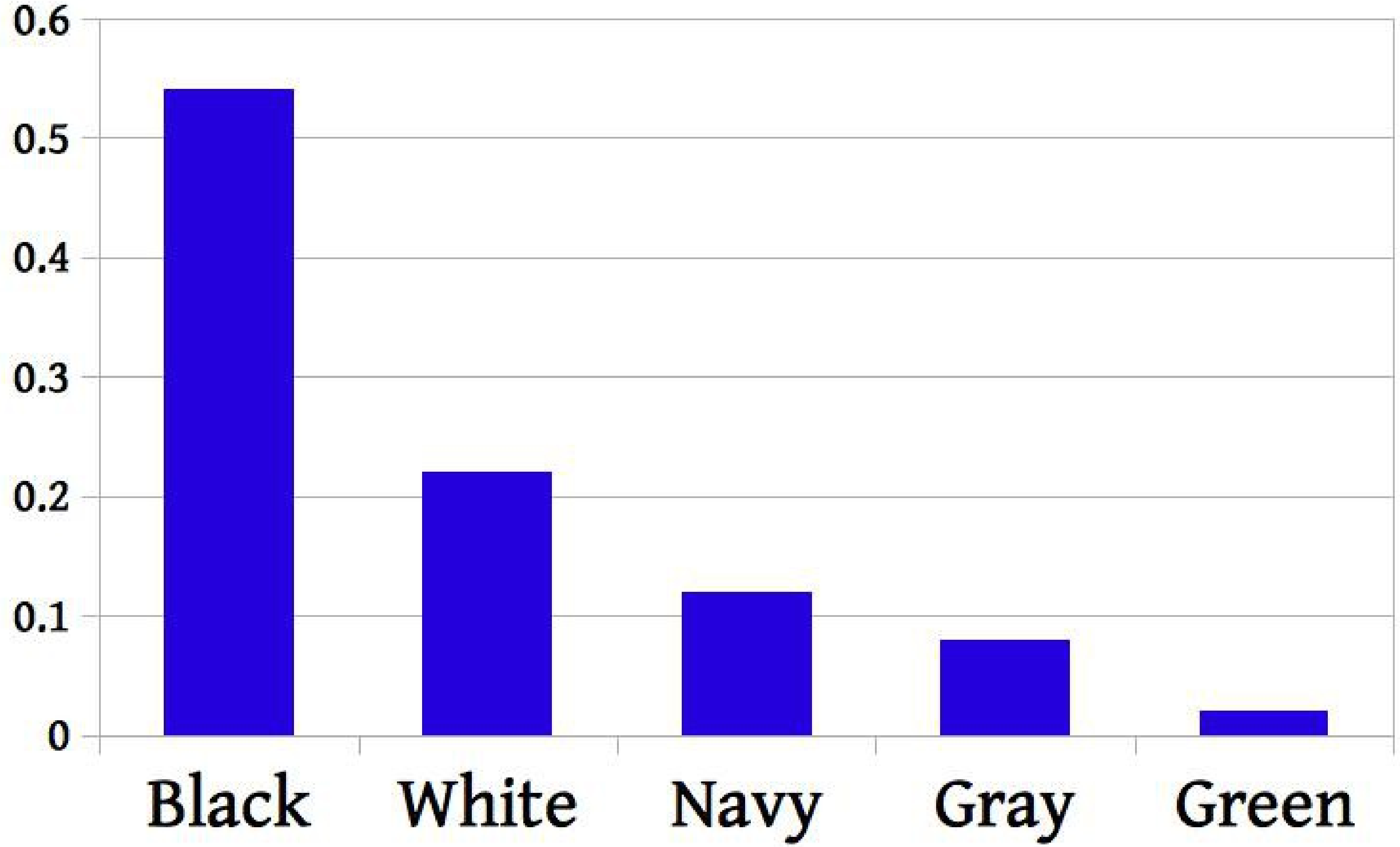
$$H(X) = E\left[\log \frac{1}{\Pr(X = x)}\right] = 2.31$$

Why not use Shannon information?



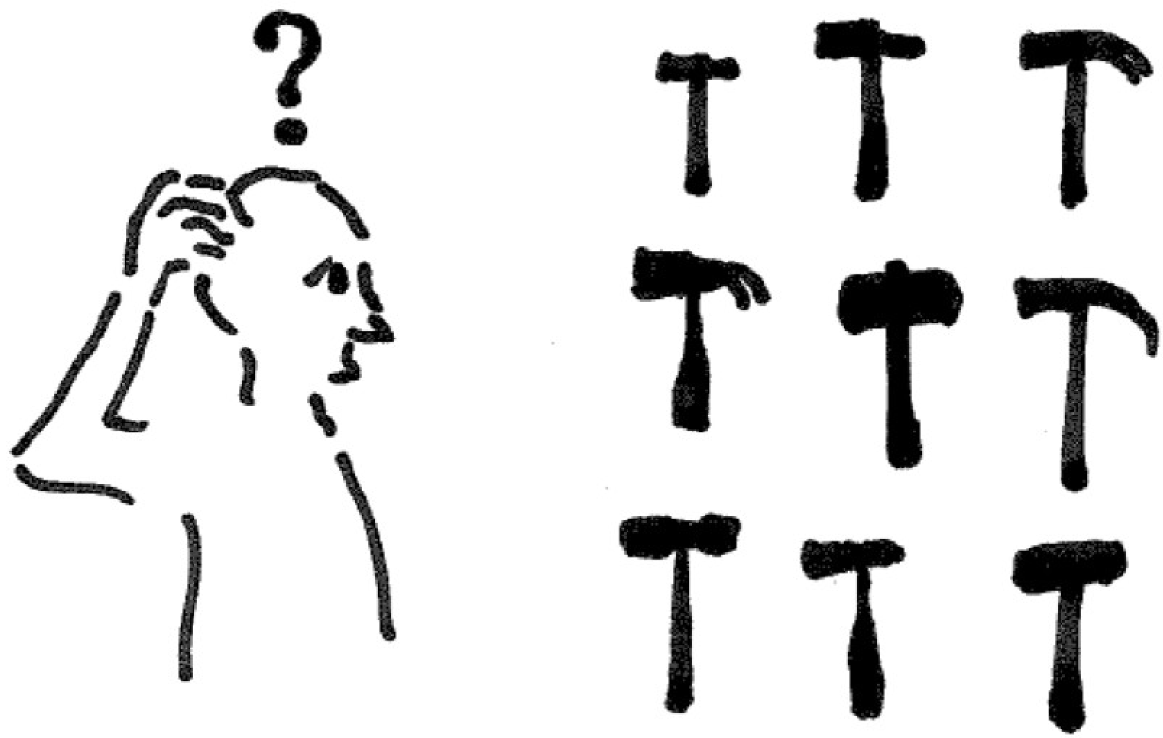
Expected number of questions: $= 2.34$

What color are my socks?

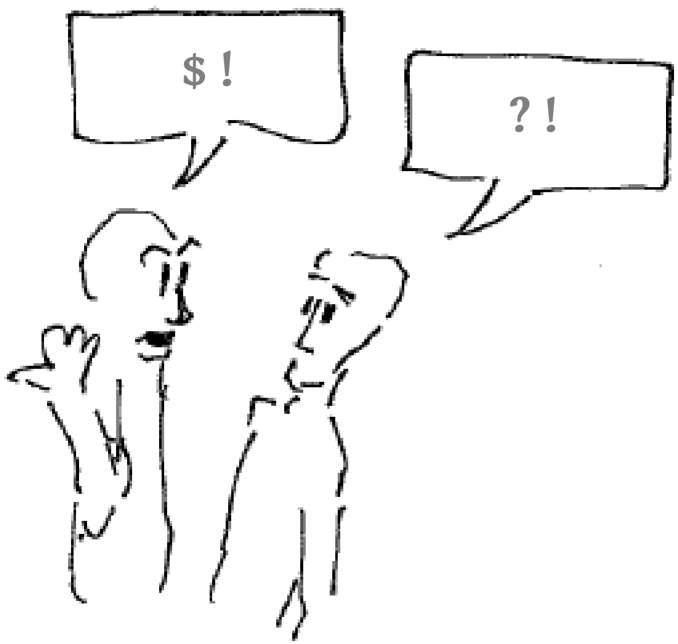


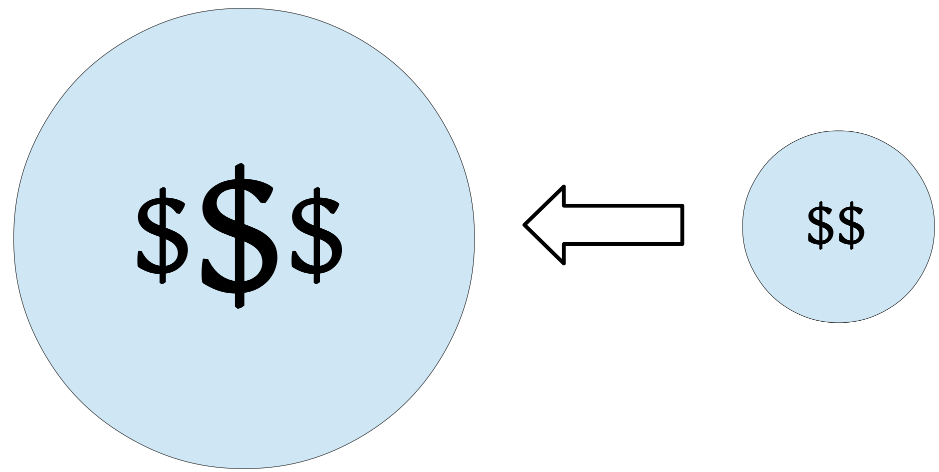
$$H(\mathbf{p}) = -\Sigma\, \mathbf{p} \log \mathbf{p} = 6.53 \text{ bits of entropy.}$$

How?



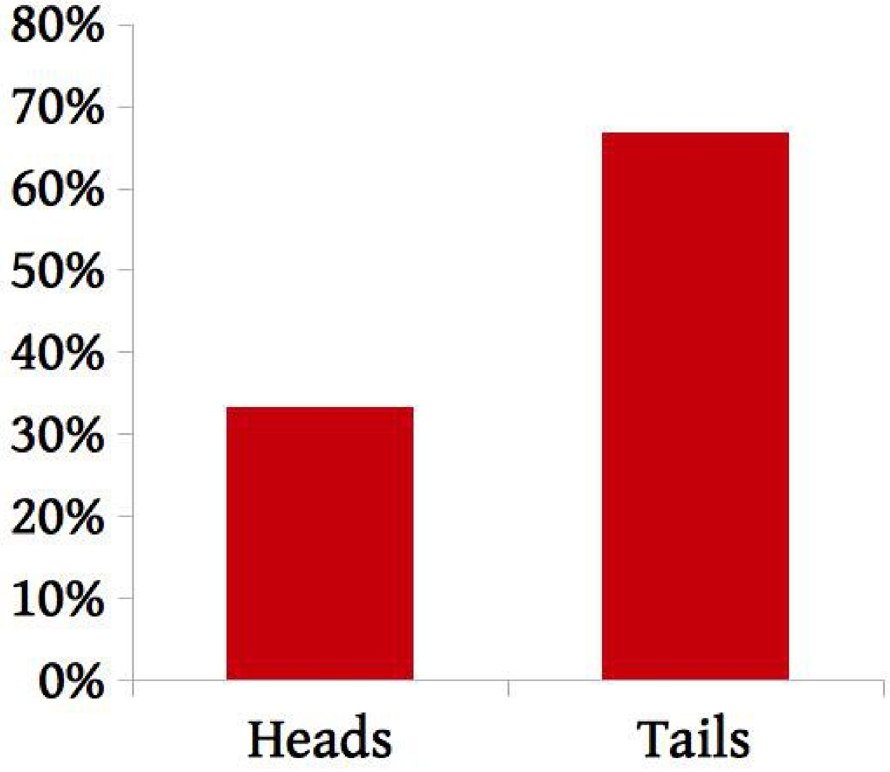
Why not use value-of-information?





Value-of-Information = Posterior Expectation − Prior Expectation

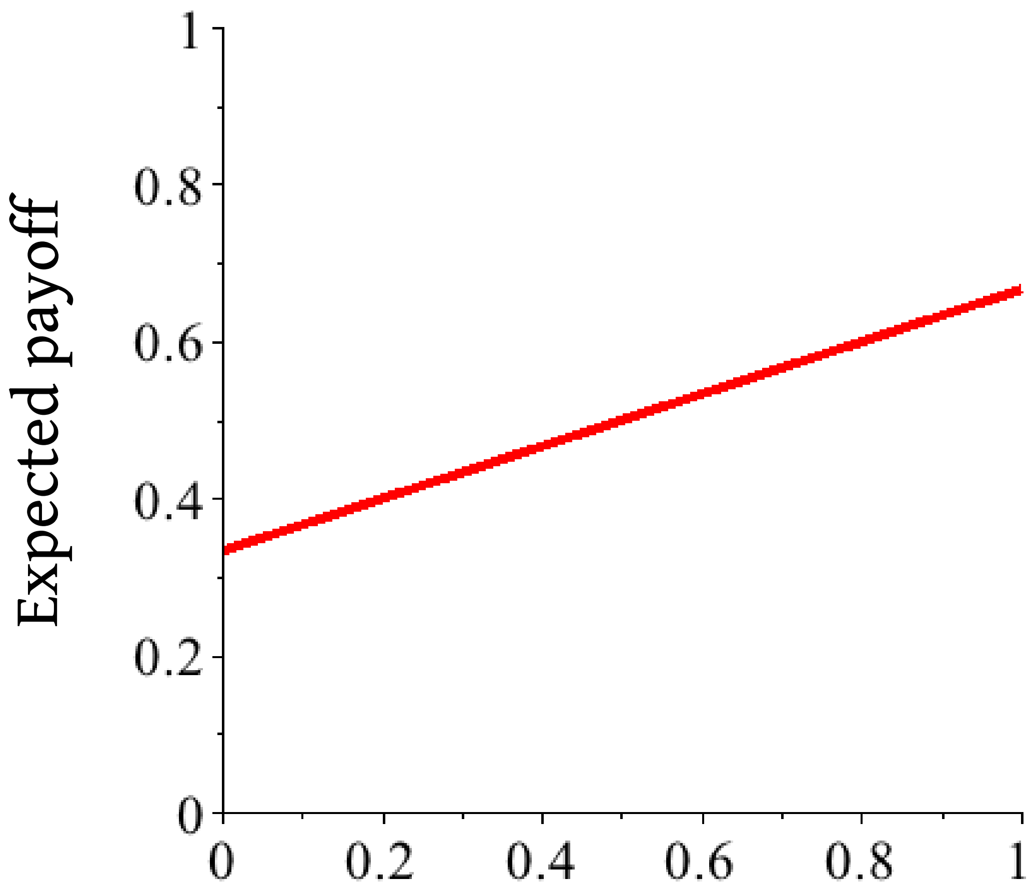
Why not use value-of-information?



<u>Rules:</u>

- Your capital can be **distributed freely**
- Bets on the **actual outcome** are returned **twofold**
- Bets on **all other outcomes** are **lost**

Why not use value-of-information?



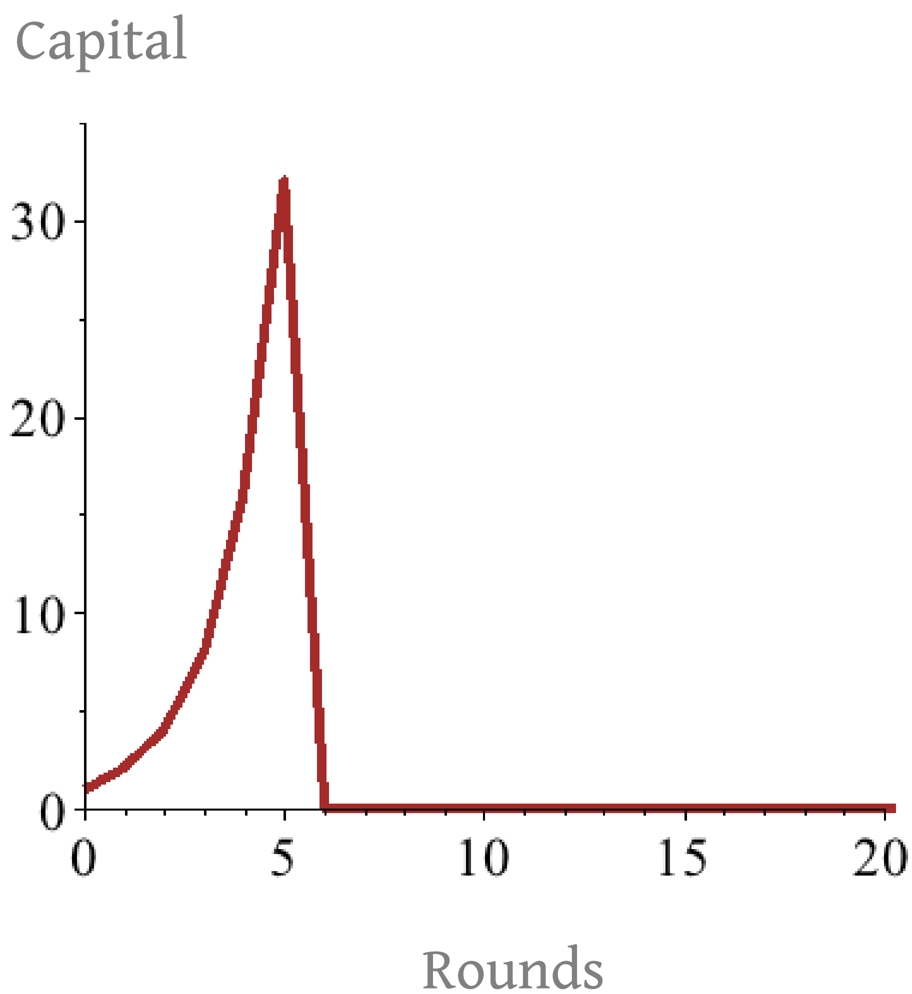
(Everything on Heads)

(Everything on Tails)

<u>Optimal Strategy:</u>

Degenerate Gambling

Why not use value-of-information?



Why not use value-of-information?

<u>Rate of return:</u>

$$R_i = \frac{\text{Capital at time } i+1}{\text{Capital at time } i}$$

<u>Long-run behavior:</u>

$$E[\, R_1 \cdot R_2 \cdot R_3 \cdots R_n \,]$$

Why not use value-of-information?

Rate of return:

$$R_i = \frac{\text{Capital at time } i+1}{\text{Capital at time } i}$$

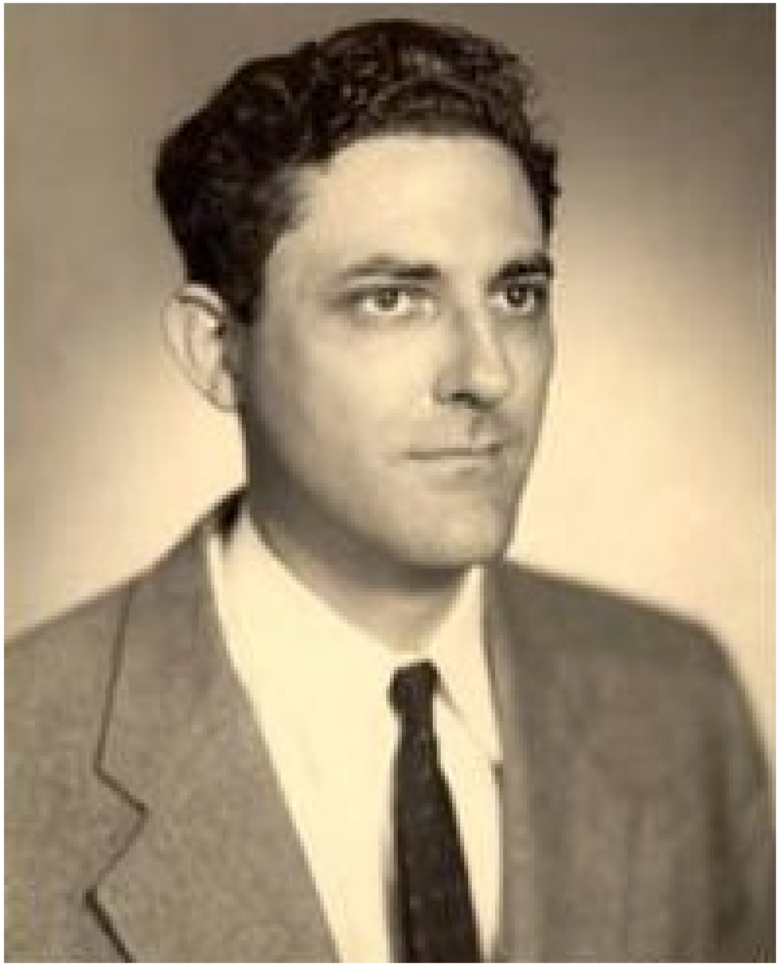
Long-run behavior:

$$E[\, R_1 \cdot R_2 \cdot R_3 \cdots R_n \,]$$

Converges to 0 in probability as n → ∞

Optimal reinvestment

Daniel Bernoulli
(1700 – 1782)



John Larry Kelly, Jr.
(1923 – 1965)

Optimal reinvestment

Doubling rate:

$$W_i = \log \frac{\text{Capital at time } i+1}{\text{Capital at time } i}$$

(so $R = 2^W$)

Optimal reinvestment

<u>Doubling rate:</u>

$$W_i = \log \frac{\text{Capital at time } i+1}{\text{Capital at time } i}$$

(so $R = 2^W$)

<u>Long-run behavior:</u>

$$E[\, R_1 \cdot R_2 \cdot R_3 \cdots R_n \,]$$

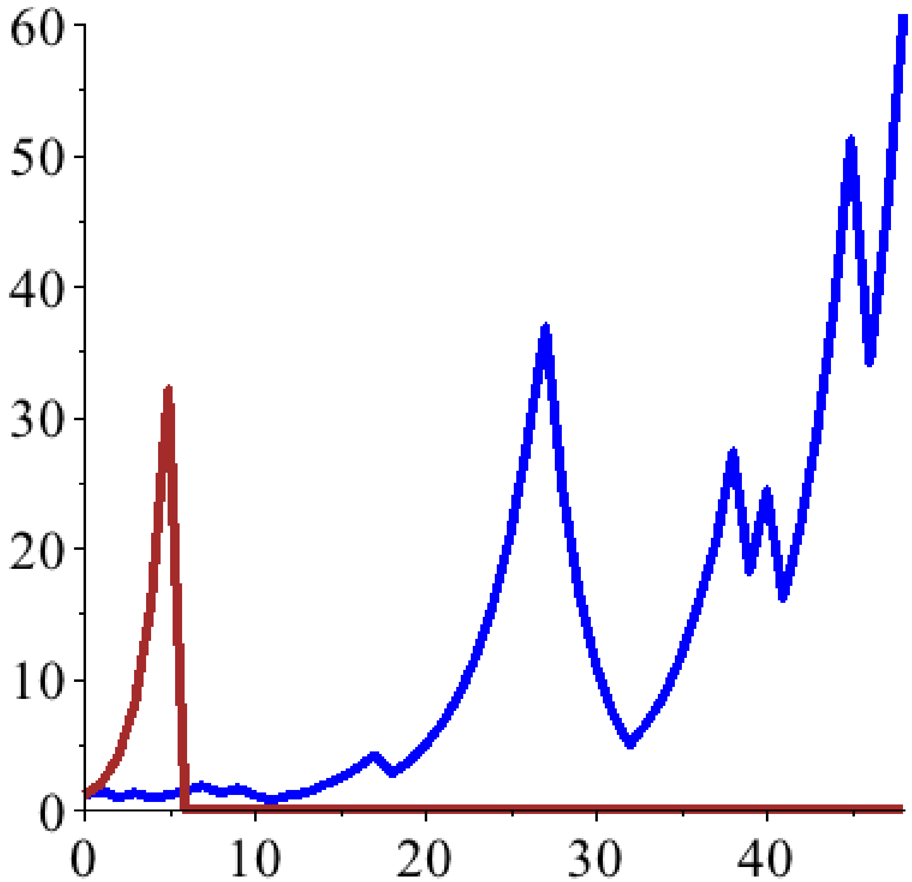
$$= E[2^{W_1 + W_2 + W_3 + \cdots + W_n}]$$

$$= 2^{E[W_1 + W_2 + W_3 + \cdots + W_n]}$$

$$\longrightarrow 2^{nE[W]}$$

for $n \to \infty$

Optimal reinvestment



Logarithmic expectation

$$E[W] = \sum \mathbf{p} \log \mathbf{bo}$$

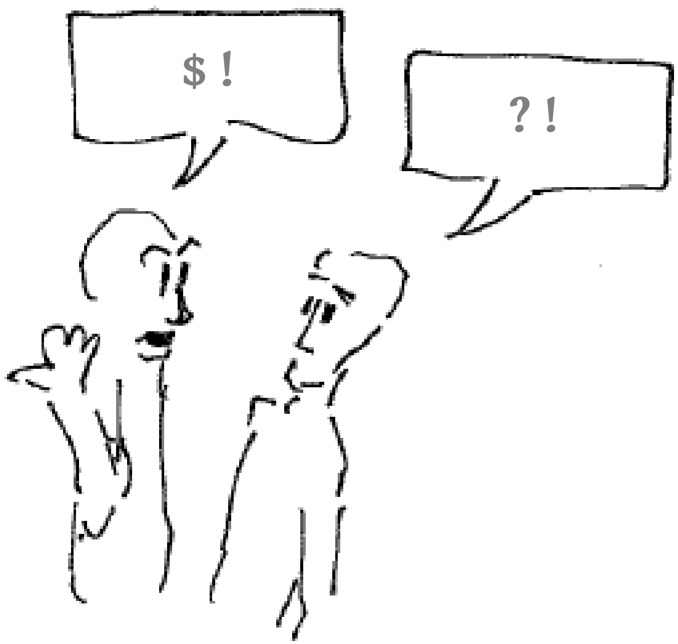
is maximized by **proportional gambling** ($\mathbf{b}^* = \mathbf{p}$).

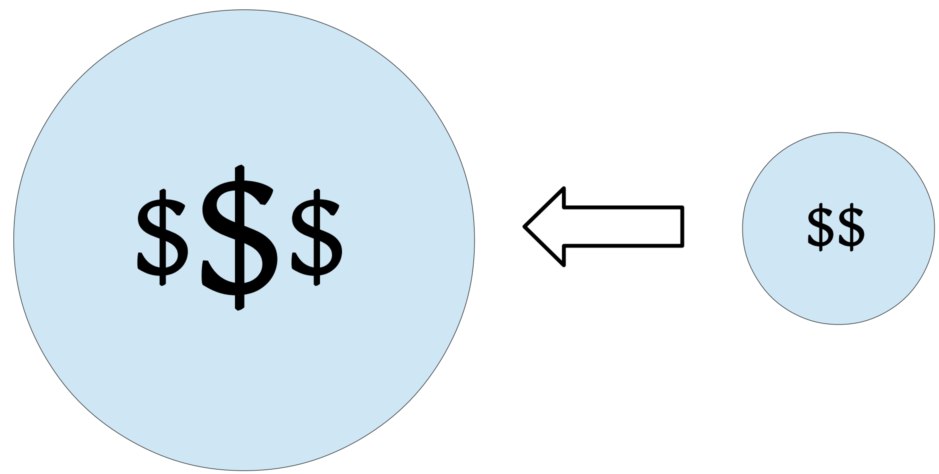
Arithmetic expectation

$$E[R] = \sum \mathbf{pbo}$$

is maximized by **degenerate gambling**

Measuring relevant information





Amount of relevant information	=	Posterior expected doubling rate	−	Prior expected doubling rate

Measuring relevant information

Definition (Relevant Information):

For an agent with utility function u, the **amount of relevant information** contained in the message $Y = y$ is

$$K(y) = \underbrace{\sum \max_s \sum \Pr(x \mid y) \log u(s, x)}_{\text{Posterior optimal doubling rate}} - \underbrace{\max_s \sum \Pr(x) \log u(s, x)}_{\text{Prior optimal doubling rate}}$$

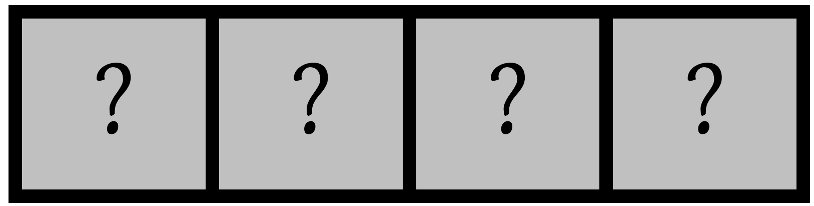
Measuring relevant information

$$K(y) = \sum \max_s \sum \Pr(x \mid y) \log u(s, x) - \max_s \sum \Pr(x) \log u(s, x)$$

- **Expected** relevant information is **non-negative.**
- Relevant information equals the **maximal fraction of future gains** you can pay for a piece of information without loss.
- When u has the form $u(s, x) = v(x)\, \mathrm{s}(x)$ for some non-negative function v, **relevant information equals Shannon information**.

Example: Code-breaking

Example: Code-breaking



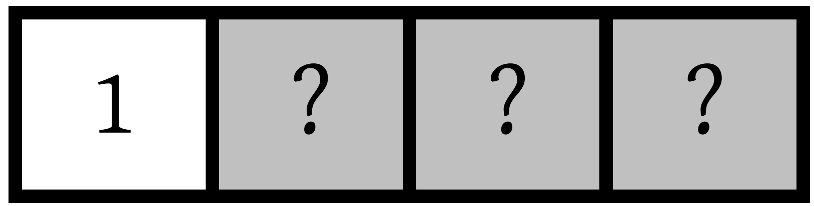
Entropy:

$H = 4$

Accumulated information:

$I(X;Y) = 0$

Example: Code-breaking



1 bit!

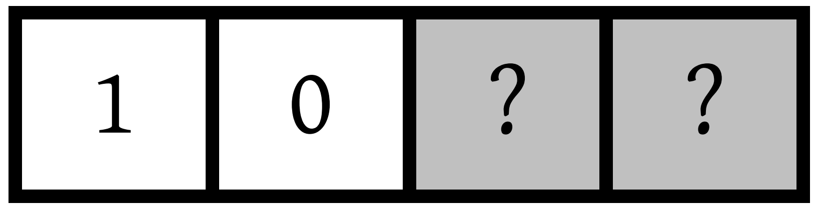
Entropy:

$H = 3$

Accumulated information:

$I(X;Y) = 1$

Example: Code-breaking



1 bit!

Entropy:

$H = 2$

Accumulated information:

$I(X; Y) = 2$

Example: Code-breaking

1	0	1	?

1 bit!

Entropy:

$H = 1$

Accumulated information:

$I(X; Y) = 3$

Example: Code-breaking

1	0	1	1

1 bit!

Entropy:

$H = 0$

Accumulated information:

$I(X;Y) = 4$

Example: Code-breaking

1	0	1	1
1 bit	**1 bit**	**1 bit**	**1 bit**

<u>Entropy:</u>

$H = 0$

<u>Accumulated information:</u>

$I(X;Y) = 4$

Example: Code-breaking

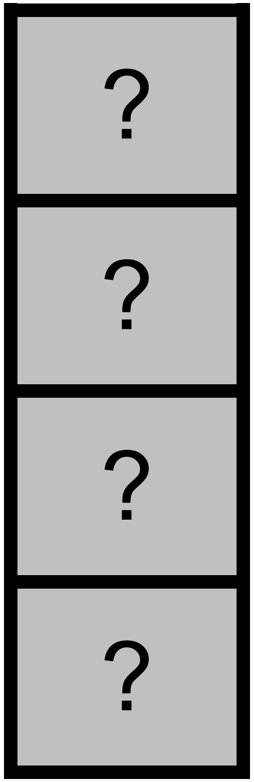
Rules:

- You can invest a fraction f of your capital in the guessing game
- If you guess the correct code, you get your investment back 16-fold:

$$u = 1 - f + 16f$$

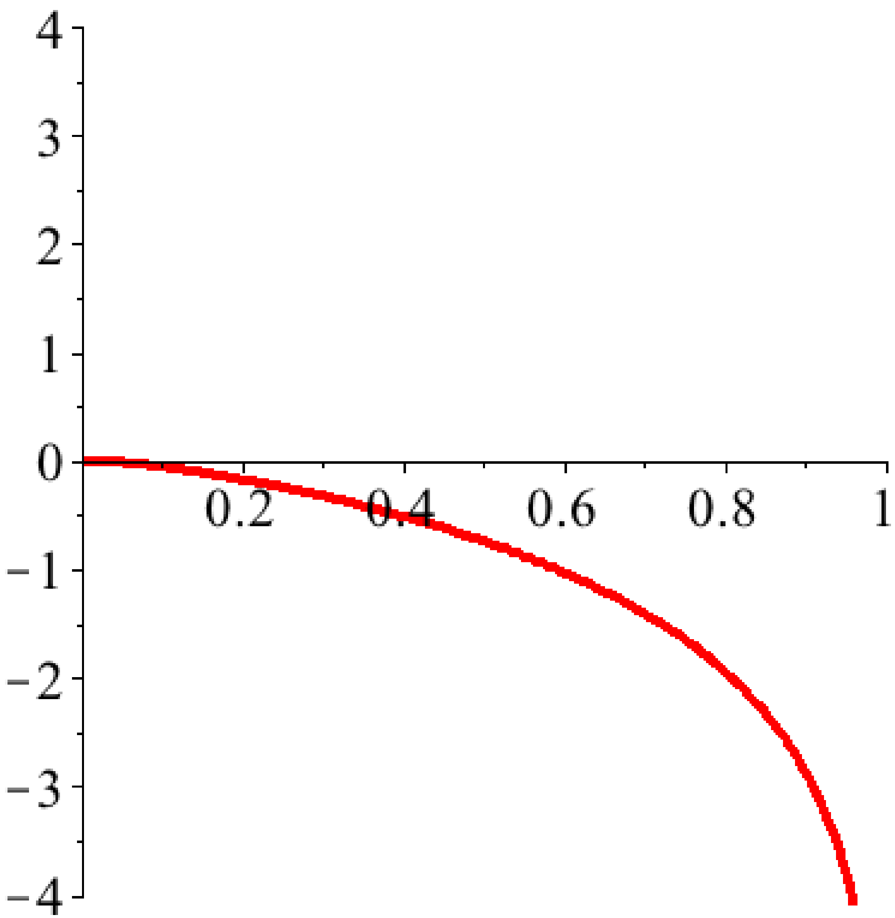
- Otherwise, you lose it:

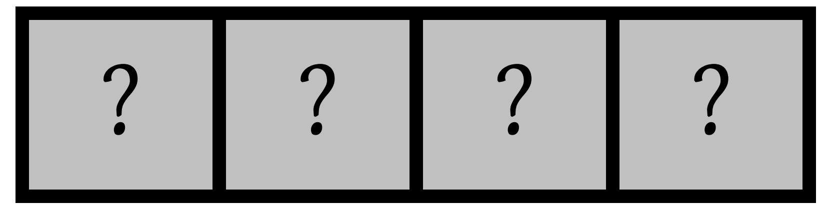
$$u = 1 - f$$



$$W(f) = \frac{15}{16}\log(1 - f) + \frac{1}{16}\log(1 - f + 16f)$$

Example: Code-breaking





<u>Optimal strategy:</u>

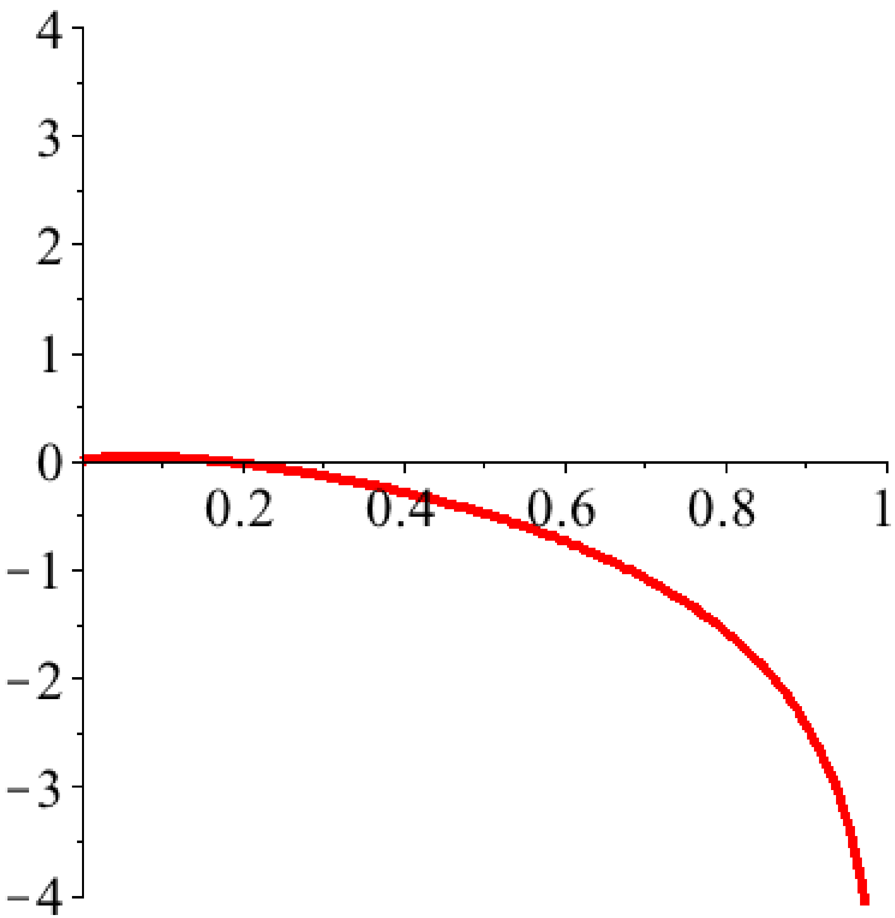
$$f^* = 0$$

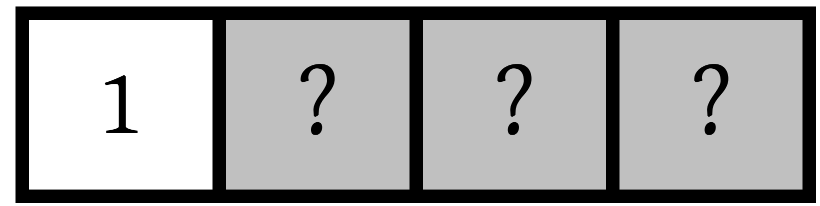
<u>Optimal doubling rate:</u>

$$W(f^*) = 0.00$$

$$W(f) = \frac{15}{16}\log(1 - f) + \frac{1}{16}\log(1 - f + 16f)$$

Example: Code-breaking





0.04 bits

Optimal strategy:

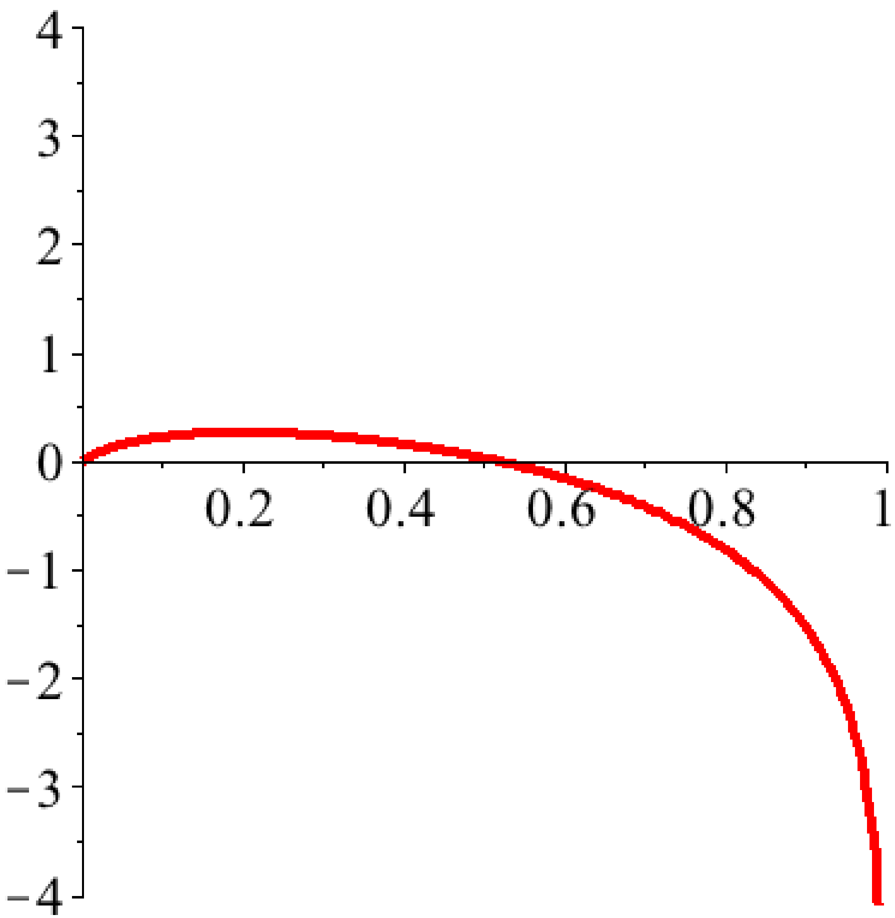
$$f^* = 1/15$$

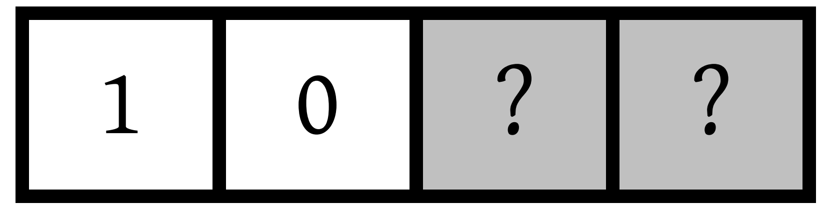
Optimal doubling rate:

$$W(f^*) = 0.04$$

$$W(f) = \frac{7}{8}\log(1-f) + \frac{1}{8}\log(1-f+16f)$$

Example: Code-breaking





0.22 bits

<u>Optimal strategy:</u>

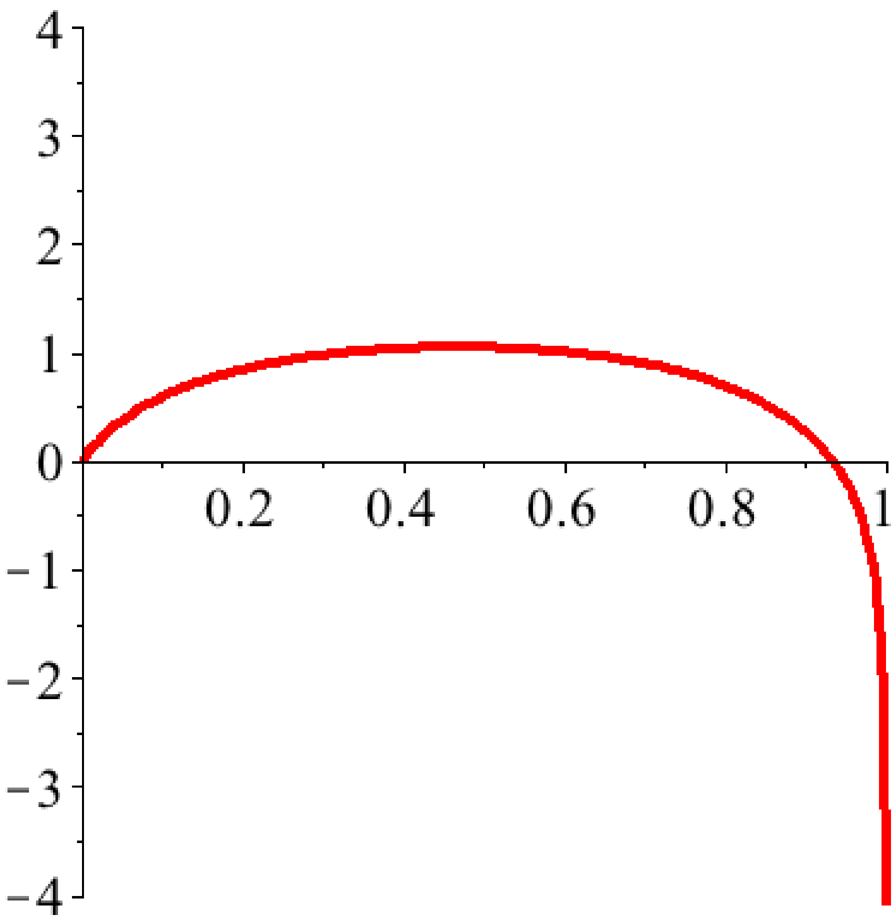
$$f^* = 3/15$$

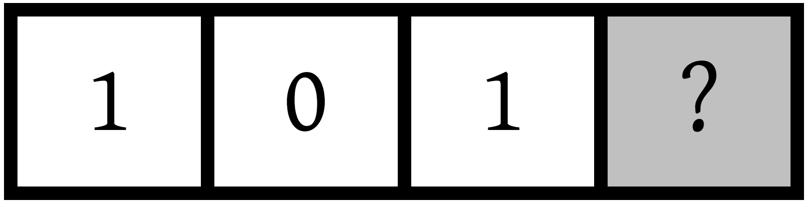
<u>Optimal doubling rate:</u>

$$W(f^*) = 0.26$$

$$W(f) = \frac{3}{4}\log(1 - f) + \frac{1}{4}\log(1 - f + 16f)$$

Example: Code-breaking





0.79 bits

<u>Optimal strategy:</u>

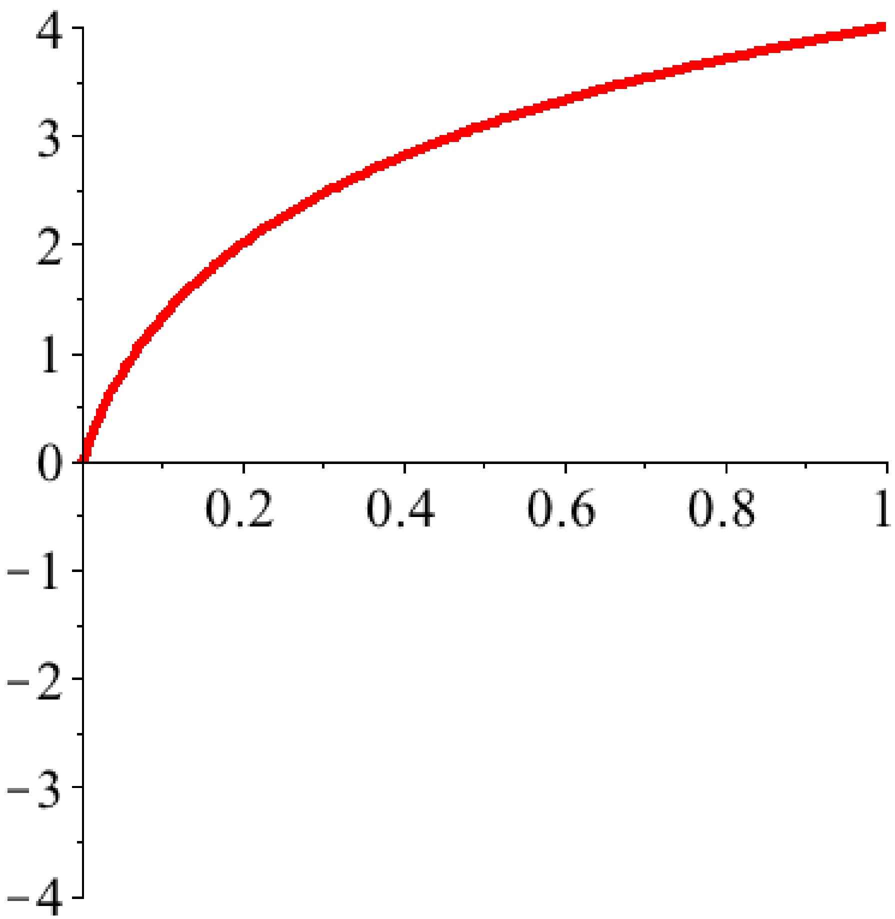
$$f^* = 7/15$$

<u>Optimal doubling rate:</u>

$$W(f^*) = 1.05$$

$$W(f) = \frac{1}{2}\log(1 - f) + \frac{1}{2}\log(1 - f + 16f)$$

Example: Code-breaking



1	0	1	1

2.95 bits

Optimal strategy:

$$f^* = 1$$

Optimal doubling rate:

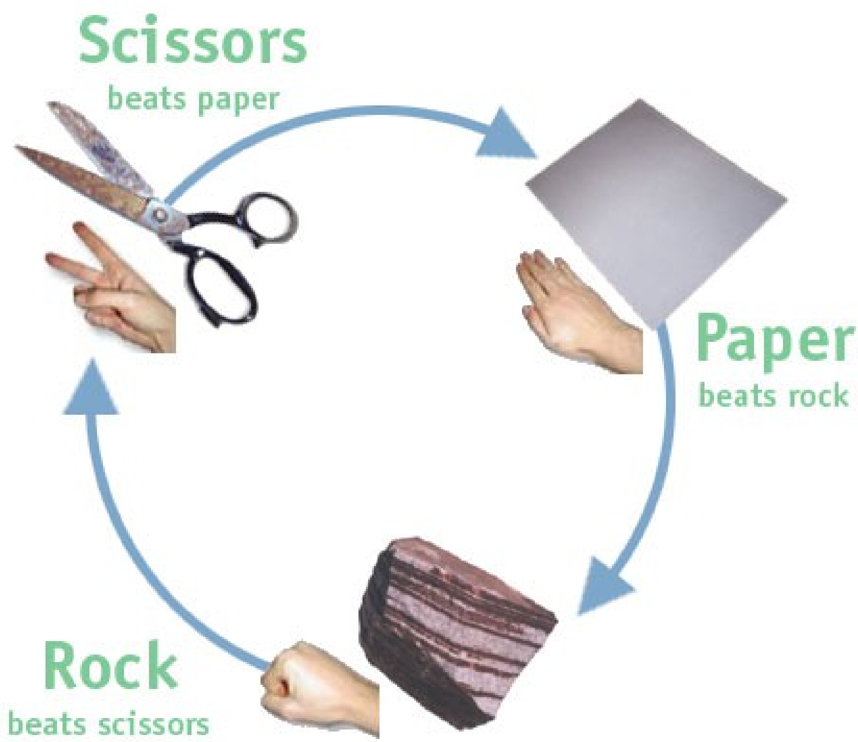
$$W(f^*) = 4.00$$

$$W(f) = \frac{0}{1}\log(1 - f) + \frac{1}{1}\log(1 - f + 16f)$$

Example: Code-breaking

	?	?	?	?
Raw information (drop in **entropy**)	**1.00**	**1.00**	**1.00**	**1.00**
Relevant information (increase in **doubling rate**)	**0.04**	**0.22**	**0.79**	**2.95**

Example: Randomization

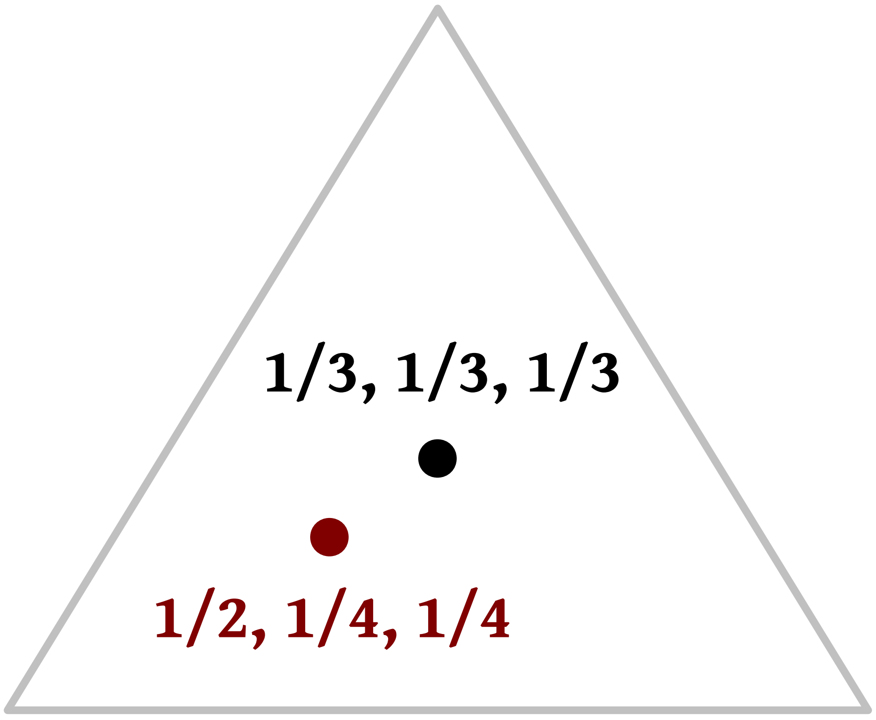


Random Number Table

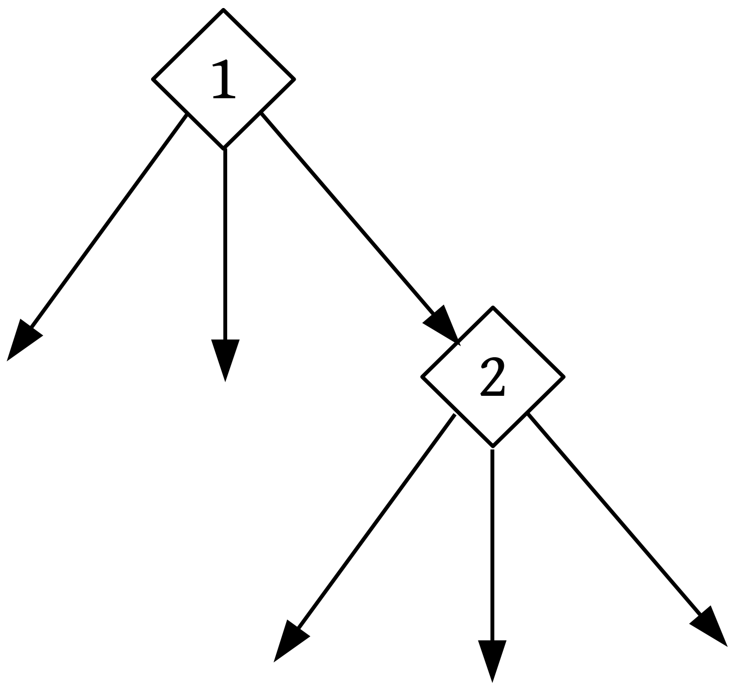
13962	70992	65172	28053	02190	83634	66012	70305	66761	88344
43905	46941	72300	11641	43548	30455	07686	31840	03261	89139
00504	48658	38051	59408	16508	82979	92002	63606	41078	86326
61274	57238	47267	35303	29066	02140	60867	39847	50968	96719
43753	21159	16239	50595	62509	61207	86816	29902	23395	72640
83503	51662	21636	68192	84294	38754	84755	34053	94582	29215
36807	71420	35804	44862	23577	79551	42003	58684	09271	68396
19110	55680	18792	41487	16614	83053	00812	16749	45347	88199
82615	86984	93290	87971	60022	35415	20852	02909	99476	45568
05621	26584	36493	63013	68181	57702	49510	75304	38724	15712
06936	37293	55875	71213	83025	46063	74665	12178	10741	58362
84981	60458	16194	92403	80951	80068	47076	23310	74899	87929
66354	88441	96191	04794	14714	64749	43097	83976	83281	72038
49602	94109	36460	62353	00721	66980	82554	90270	12312	56299
78430	72391	96973	70437	97803	78683	04670	70667	58912	21883
33331	51803	15934	75807	46561	80188	78984	29317	27971	16440
62843	84445	56652	91797	45284	25842	96246	73504	21631	81223
19528	15445	77764	33446	41204	70067	33354	70680	66664	75486
16737	01887	50934	43306	75190	86997	56561	79018	34273	25196
99389	06685	45945	62000	76228	60645	87750	46329	46544	95665
36160	38196	77705	28891	12106	56281	86222	66116	39626	06080
05505	45420	44016	79662	92069	27628	50002	32540	19848	27319
85962	19758	92795	00458	71289	05884	37963	23322	73243	98185
28763	04900	54460	22083	89279	43492	00066	40857	86568	49336
42222	40446	82240	79159	44168	38213	46839	26598	29983	67645
43626	40039	51492	36488	70280	24218	14596	04744	89336	35630
97761	43444	95895	24102	07006	71923	04800	32062	41425	66862

Example: Randomization

```
def choose():

   if flip():

      if flip():
         return ROCK

      else:
         return PAPER

   else:
      return SCISSORS
```



Example: Randomization

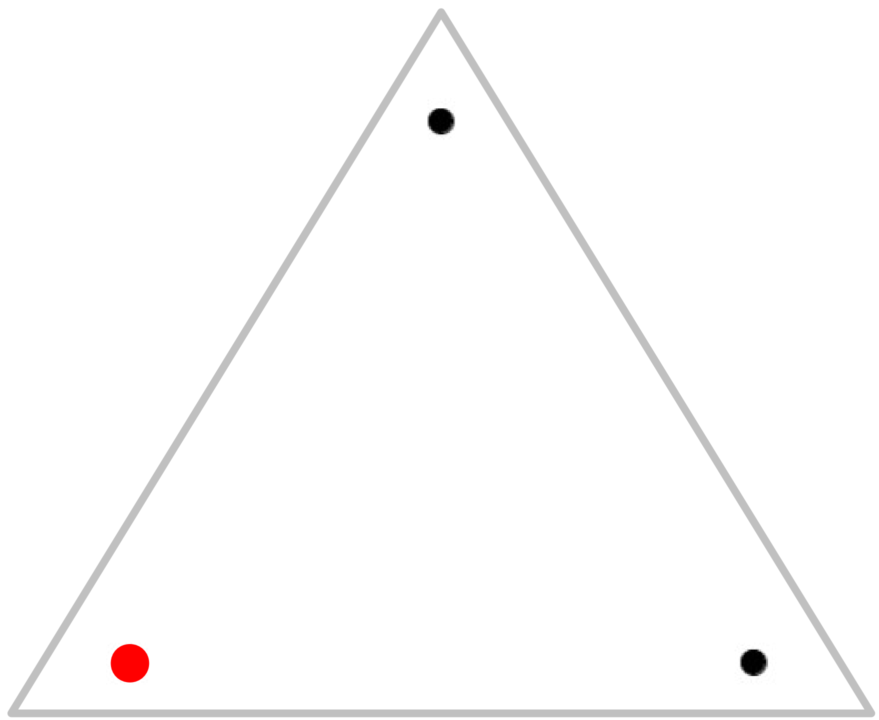


Rules:

- You (1) and the adversary (2) both bet $1
- You move first
- The winner takes the whole pool

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization



Best accessible strategy:

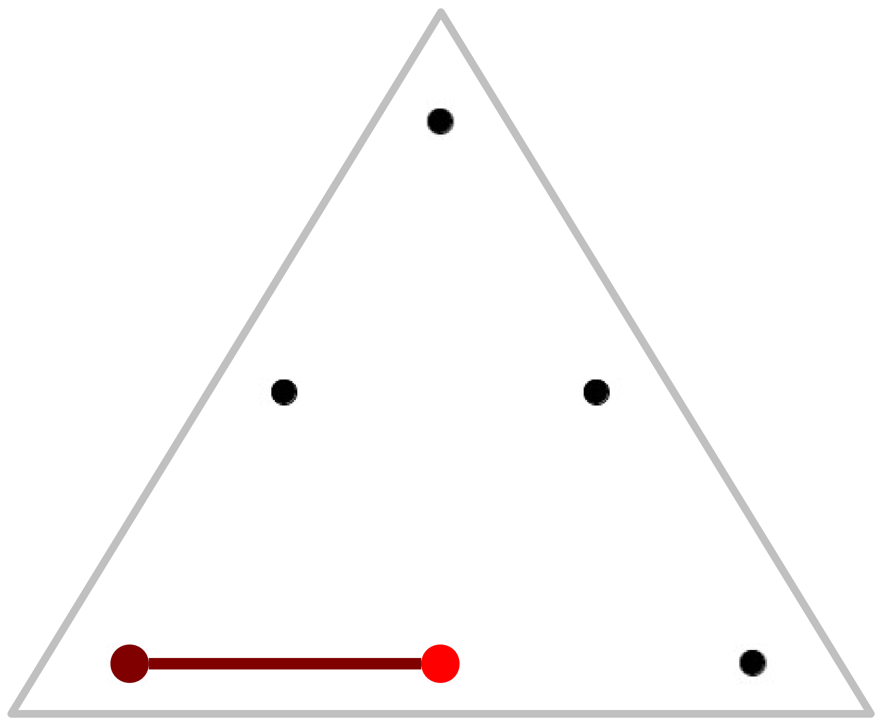
$$\mathbf{p}^* = (1, 0, 0)$$

Doubling rate:

$$W(\mathbf{p}^*) = -\infty$$

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization



Best accessible strategy:

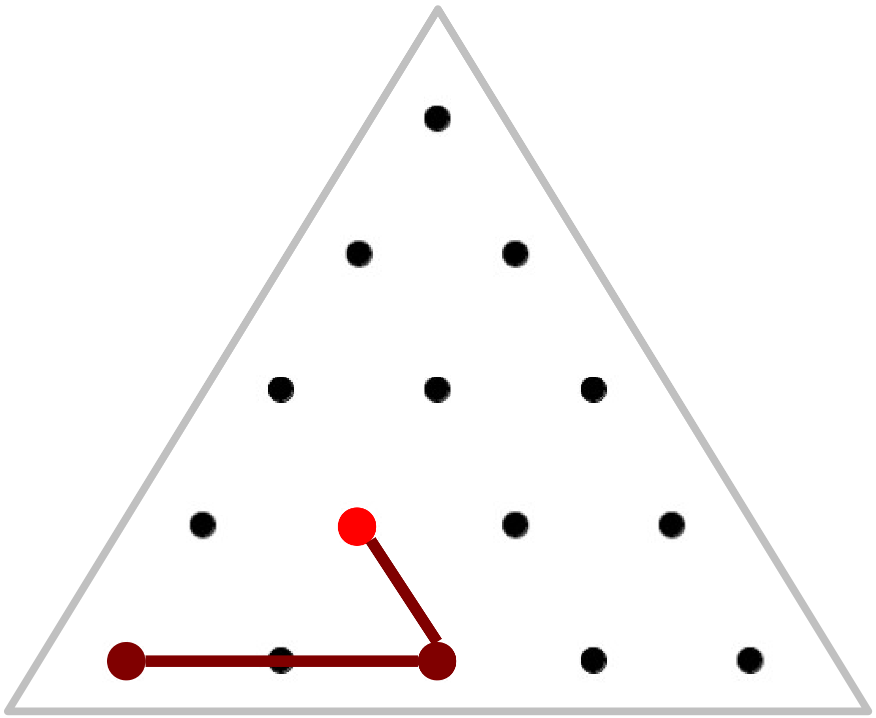
$$\mathbf{p}^* = (1/2, 1/2, 0)$$

Doubling rate:

$$W(\mathbf{p}^*) = -1.00$$

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization



Best accessible strategy:

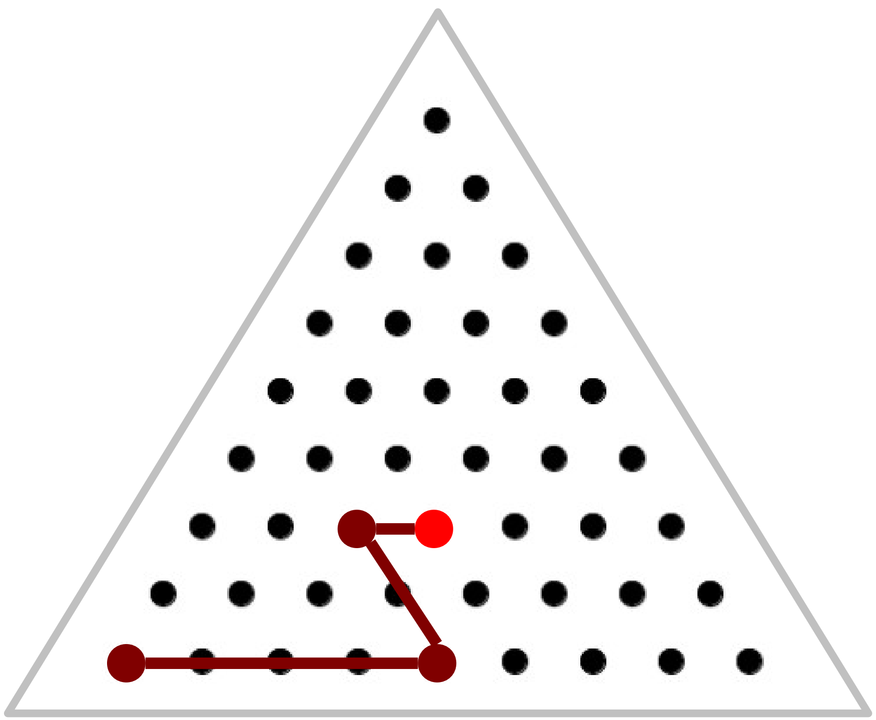
$$\mathbf{p}^* = (2/4, 1/4, 1/4)$$

Doubling rate:

$$W(\mathbf{p}^*) = -0.42$$

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization



Best accessible strategy:

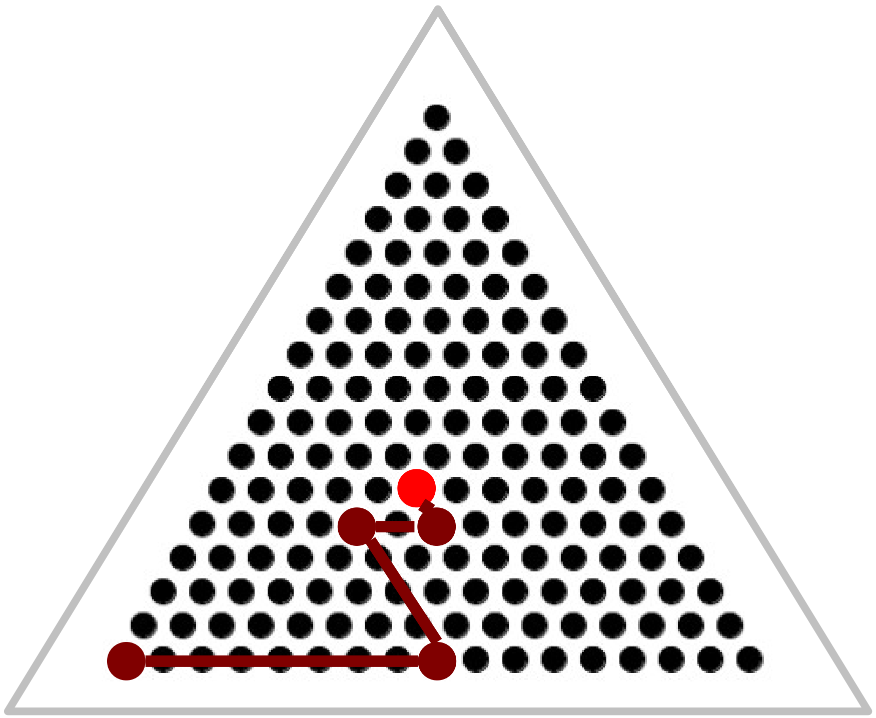
$$\mathbf{p}^* = (3/8, 3/8, 2/8)$$

Doubling rate:

$$W(\mathbf{p}^*) = -0.19$$

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization



Best accessible strategy:

$$\mathbf{p}^* = (6/16, 5/16, 5/16)$$

Doubling rate:

$$W(\mathbf{p}^*) = -0.09$$

$$W(\mathbf{p}) = \log \min \{ \mathbf{p}_1 + 2\,\mathbf{p}_2,\ \mathbf{p}_2 + 2\,\mathbf{p}_3,\ \mathbf{p}_3 + 2\,\mathbf{p}_1 \}$$

Example: Randomization

Coin flips	Distribution	Doubling rate
0	(1, 0, 0)	$-\infty$
1	(1/2, 1/2, 0)	−1.00
2	(1/2, 1/4, 1/4)	−0.42
3	(3/8, 3/8, 2/8)	−0.19
4	(6/16, 5/16, 5/16)	−0.09
...	...	...
∞	(1/3, 1/3, 1/3)	0.00

∞

0.58

0.23

0.10

A Quantitative Measure of Relevance Based on Kelly Gambling Theory

Mathias Winther Madsen

ILLC, University of Amsterdam

Defining a good concept of relevance is a key problem in all disciplines that theorize about information, including information retrieval [3], epistemology [5], and the pragmatics of natural languages [12].

Shannon information theory [10] provides an interesting quantification of the notion of information, but it does not in itself provide any tools for distinguishing useless from useful facts. The microeconomic concept of value-of-information [1] does provide tools for doing so, but it is not easily combined with information theory, and is largely unable to exploit any of its tools or insights.

In this paper, I propose a framework that integrates information theory more natively with utility theory and thus tackles these problems. Specifically, I draw on John Kelly's application of information theory to gambling situations [7]. Kelly showed that when we take logarithmic capital growth as our measure of real utility, information theory can integrate seamlessly with classical Bernoulli gambling theory. My approach here is to turn this approach on its head and base a notion of information directly on the concept of utility.

The resulting measure coincides with Shannon information in situations in which any piece of information can be converted into a strategy improvement. When the environment provides both useful and useless information, the concept explains and quantifies the difference, and thus suggests a novel notion of value-of-information.

1 Doubling Rates and Kelly Gambling

In real gambling situations, people will often evaluate a strategy in terms of its effect on the **growth rate** of their capital, that is,

$$R = \frac{\text{Posterior capital}}{\text{Prior capital}}.$$

However, using growth rate as your utility measure suggests a gambling strategy in which you bet your entire capital on the single most likely event. Such a strategy assigns a non-zero probability to the event of losing the whole capital. If it is used in repeated plays of the same game, it thus leads to eventual bankruptcy with probability 1.

If we are instead interested in maximizing the long-term growth of a stock of capital through repeated investment and reinvestment, a better measure of strategy quality is the logarithm of the growth rate,

$$W = \log R.$$

When the logarithm is base two, this quantity is called the **doubling rate** of the capital, in analogy with the half-life of a radioactive material. W measures how many times your capital is expected to be doubled in a single game, and $1/W$ the average number of games it takes to double your capital once.

Because $\log 0^+ = -\infty$, your doubling rate will be $-\infty$ if there is even the slightest chance that you lose your whole capital by the strategy you are using. Consequently, using the doubling rate as your measure of utility will discourage strategies that can lead to bankruptcy and instead lead to a strategy that maximizes long-term exponential growth [4, ch. 6].

For the purposes of the present paper, however, the doubling rate is also interesting because of its seamless integration with Shannon information theory. To see this, consider a horse race in which horse x has probability $p(x)$ of winning. By the method of Lagrange multipliers, we can find that independently of the the odds on the horses, a gambler's doubling rate is maximized by **proportional betting** [4], i.e., betting a fraction of $p(x)$ of the total capital on horse x. If you know what these probabilities are, you thus know what the optimal strategy is.

In general, however, a gambler may have have only little or bad information about the horses, and thus use an inferior probability estimate $q \approx p$. Using the probability distribution q as a capital distribution scheme, the gambler's doubling rate will then be as follows, assuming that the odds are expressed as $c/r(x)$ for some constant c and some positive function r:

$$\begin{aligned} W(q) &= \sum_x p(x) \log\left(c \times \frac{q(x)}{r(x)}\right) \\ &= \sum_x p(x) \log\left(c \times \frac{p(x)}{r(x)} \times \frac{q(x)}{p(x)}\right) \\ &= \sum_x p(x) \log\left(\frac{p(x)}{r(x)}\right) - \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) + \log c. \end{aligned}$$

The second term in this expression, $\sum_x p(x) \log(p(x)/r(x))$, is the **Kullback-Leibler divergence** [8] between p and r, and is also written $D(p\,||\,r)$. It is an measure of how big an error the probability estimate r induces in an environment with actual probabilities p. Similarly, the second term is $D(p\,||\,q)$, the divergence from p to q.

It thus turns out that the bookmaker and the gambler are in a symmetric situation: Both the distribution of bets (q) and the size of the odds (r) implicitly express subjective probability estimates. The payoffs for the gambler and the bookmaker are determined by the quality of these estimates.

In particular, if $c = 1$, the player with the probability estimate closest to p in informational terms will make money at the expense of the other. Further, if one of the two players acquire 1 bit of information about the real winner of the race, this signal can be converted into an increase of 1 capital doubling per game. In the horse race model, information thus translates directly into utility.

However, this correspondence rests on assumptions that are particular to the horse race model, including the fact that the situation involves only one random

January: Project course in information theory

Day 1: Uncertainty and Inference

Probability theory:
Semantics and expressivity
Random variables
Generative Bayesian models
stochastic processes

Uncertain and information:
Uncertainty as cost
The Hartley measure
Shannon information content and entropy
Huffman coding

Day 2: Counting Typical Sequences

The law of large numbers
Typical sequences and the source coding theorem.

Stochastic processes and entropy rates
the source coding theorem for stochastic processes
Examples

Day 3: Guessing and Gambling

Evidence, likelihood ratios, competitive prediction
Kullback-Leibler divergence
Examples of diverging stochastic models
Expressivity and the bias/variance tradeoffs.

Doubling rates and proportional betting
Card color prediction

Day 4: Asking Questions and Engineering Answers

Questions and answers (or experiments and observations)
mutual information
Coin weighing
The maximum entropy principle

The channel coding theorem

Day 5: Informative Descriptions and Residual Randomness

The practical problem of source coding
Kraft's inequality and prefix codes
Arithmetic coding

Kolmogorov complexity
Tests of randomness
Asymptotic equivalence of complexity and entropy

Information Theory

Studiegidsnummer	5314INTH6Y
Admin. code	OWII
Studielast	6
Periode(n)	Semester 2 block 1, voertaal English
Onderwijsinstituut	Graduate School of Informatics
Docent(en)	C. Schaffner (coördinator)
Onderdeel van	Master's in Logic Computer Science (joint progr with VU University A'dam)

Voeg vak toe aan planner Vakaanmelden (alleen Geesteswetenschappen)

Leerdoelen
Understand basic concepts of Shannon's information theory (http://en.wikipedia.org/wiki/Information_theory)

Inhoud
Information theory was developed by Claude E. Shannon in the 1950s to investigate the fundamental limits on signal-processing operations such as compressing data and on reliably storing and communicating data. These tasks have turned out to be fundamental for all of computer science.

In this course, we quickly review the basics of probability theory and introduce concepts such as (conditional) Shannon entropy, mutual information and Renyi entropy. Then, we prove Shannon's theorems about data compression and channel coding. Later in the course, we also cover some aspects of information-theoretic security such as the concept of randomness extraction and privacy amplification.

Aanmelden
Registration is required via https://www.sis.uva.nl until four weeks before the start of the semester.

Onderwijsvorm
This is a 6 ECTS course, which comes to roughly 20 hours of work per week.

There will be homework exercises every week to be handed in one week later before the start of the exercise session on Friday. The answers should be in English (feel free to use LaTeX, but readable handwritten solutions are fine). Cooperation while solving the exercises is allowed and encouraged, but everyone has to hand in their own solution set in their own words.