Karlsruhe Institute of Technology

Workshop Information Theory as a Bridge Across the Geosciences and Modeling Sciences

Uwe Ehret (KIT)

Introduction to Information Theory

Overview

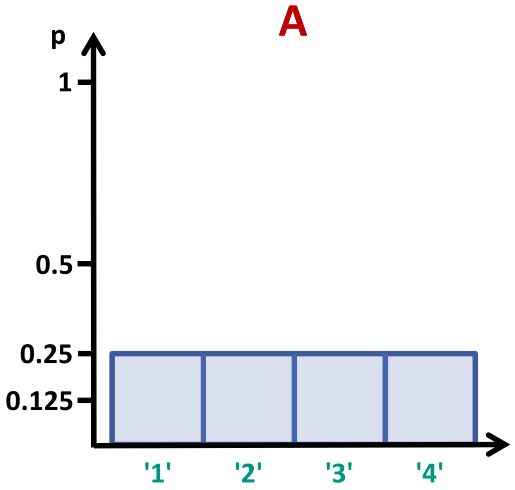
- Introduction
- Basics
- Limits
- Independence, Redundancy, Synergy
- Conservation, Dissipation, Innovation
- Applications
- Further reading

Introduction

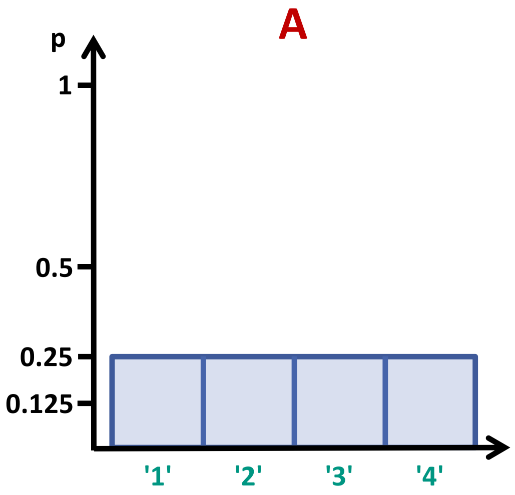
Introduction

Data ≠ Information

Introduction



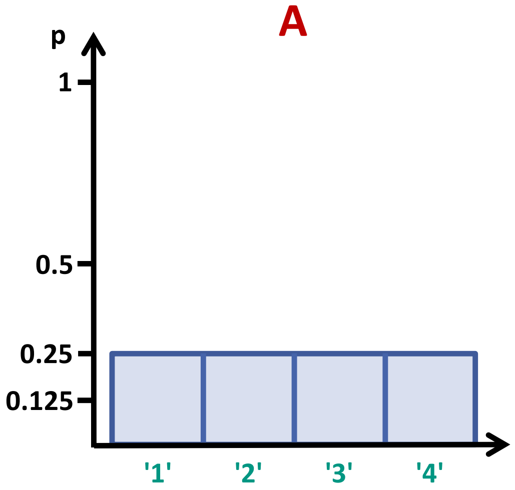
Introduction

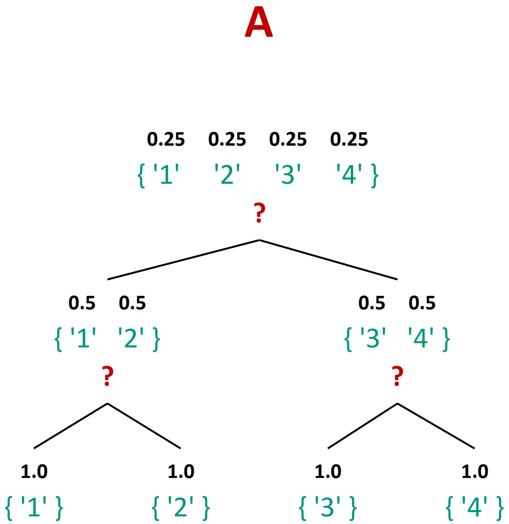


A

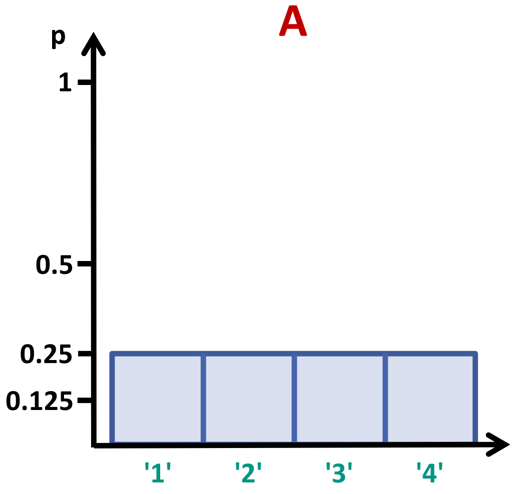
$$\{ \overset{0.25}{\text{'1'}}\quad \overset{0.25}{\text{'2'}}\quad \overset{0.25}{\text{'3'}}\quad \overset{0.25}{\text{'4'}} \}$$

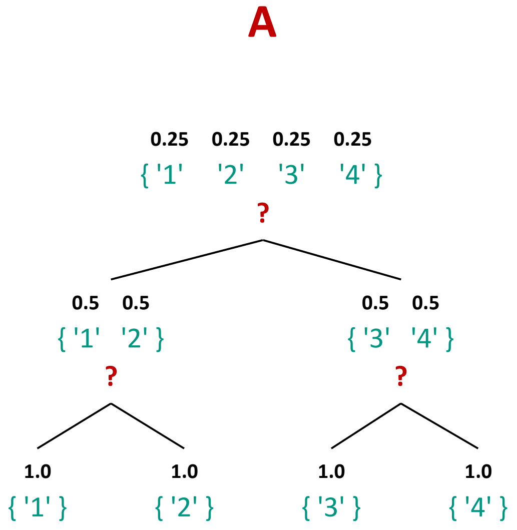
Introduction





Introduction



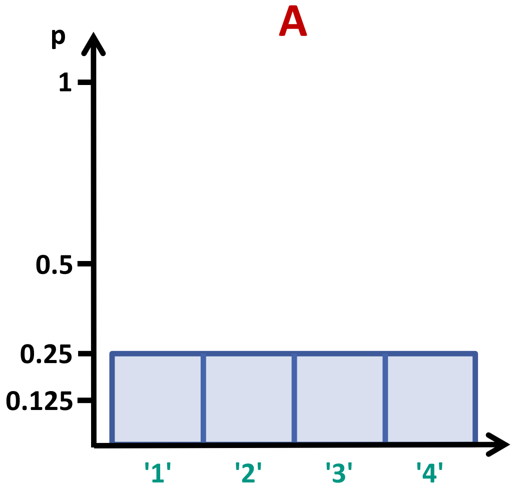


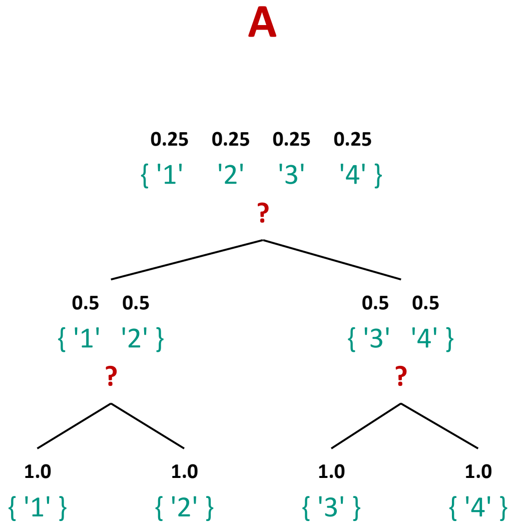
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25

Basics

- Information

$$i(x) = -log_2\big(p(x)\big)$$

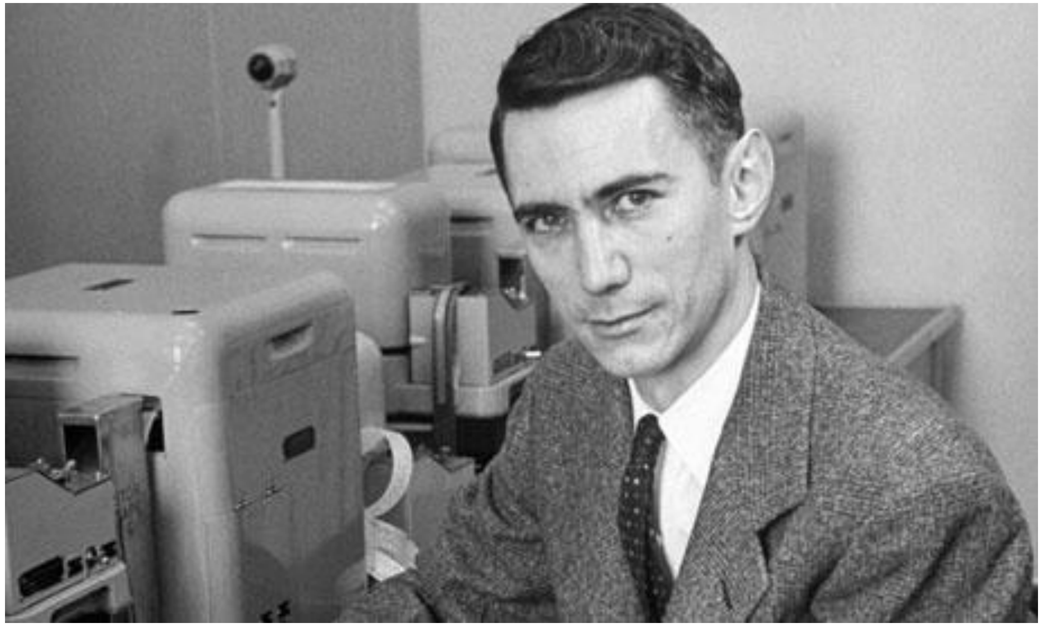




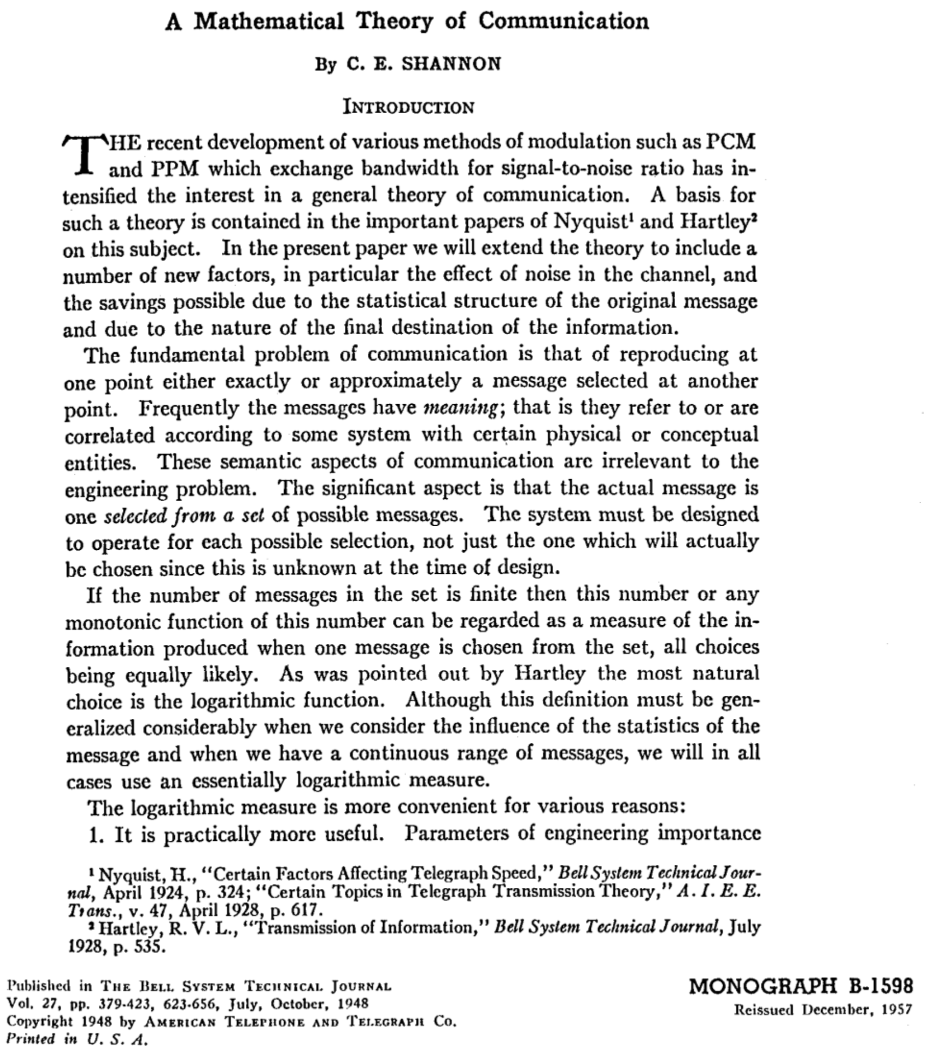
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

History

- The "Big Bang" of Information Theory



$$i(x) = -log_2\big(p(x)\big)$$



A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

Published in THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 27, pp. 379-423, 623-656, July, October, 1948
Copyright 1948 by AMERICAN TELEPHONE AND TELEGRAPH CO.
Printed in U. S. A.

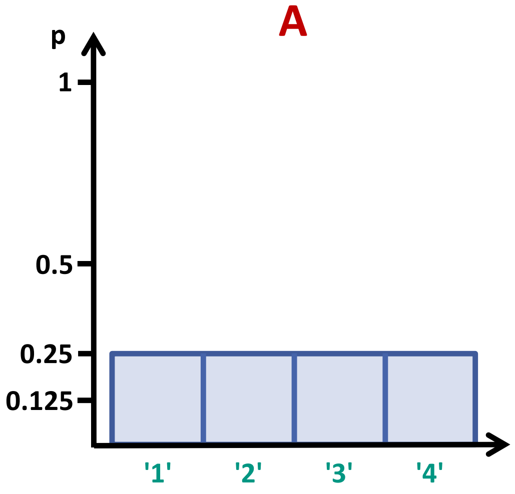
MONOGRAPH B-1598
Reissued December, 1957

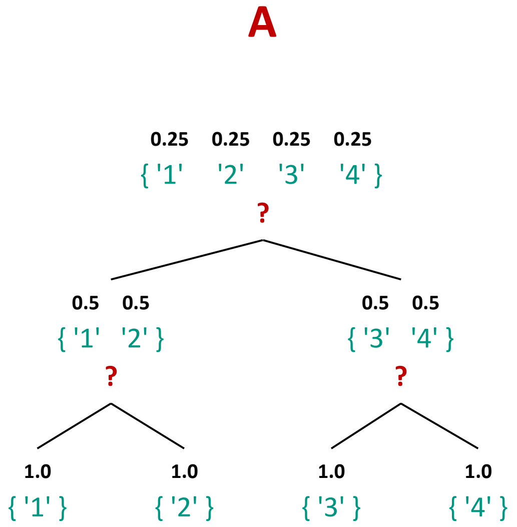
Shannon, C. E. (1948): A mathematical theory of communication, Bell system technical journal, 27, 623-656
Bild: www.theguardian.com

Basics

- Information

$$i(x) = -log_2\big(p(x)\big)$$





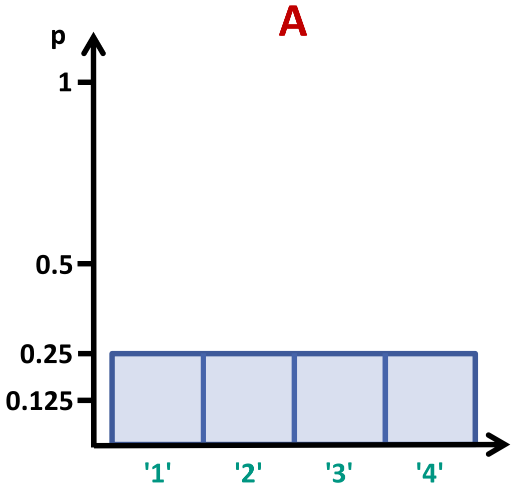
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

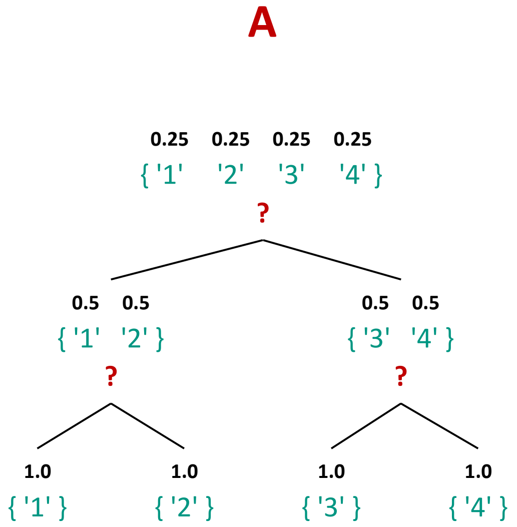
Information = # of optimal binary questions between prior and posterior state of uncertainty
1 question = 1 bit

Basics

- Information

$$i(x) = -log_2\big(p(x)\big)$$





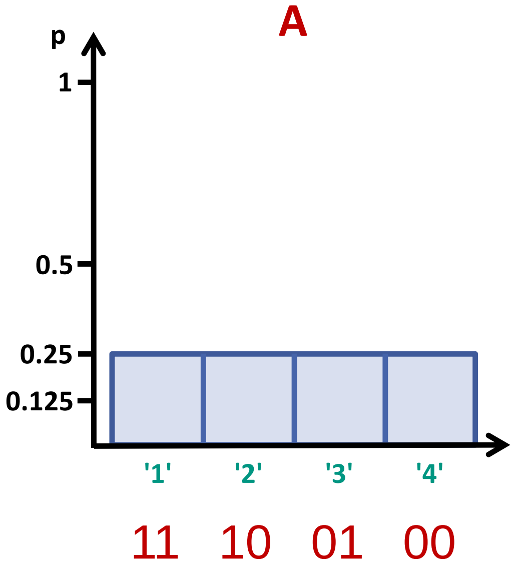
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

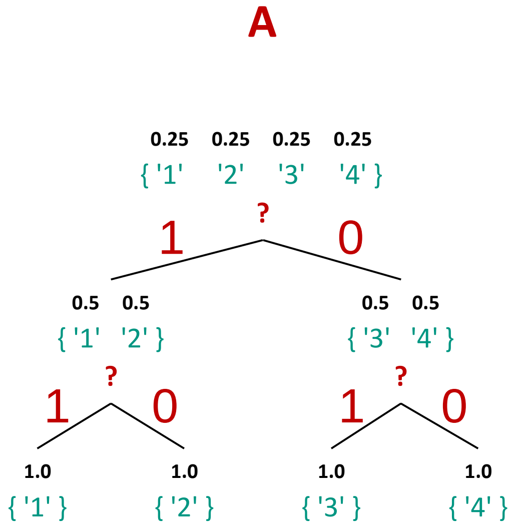
Uncertainty = missing information

Basics

- Information

$$i(x) = -log_2\big(p(x)\big)$$



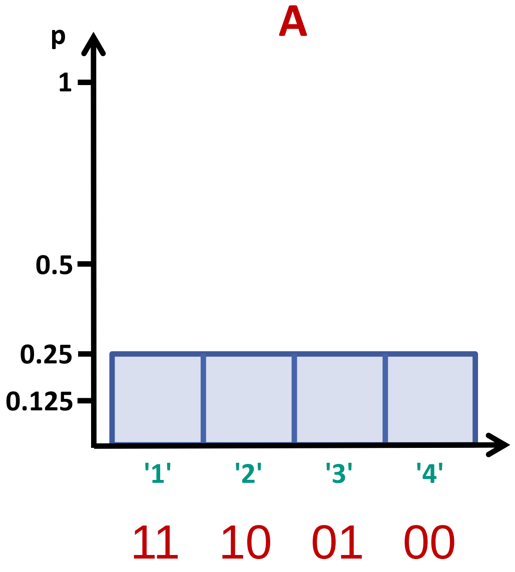


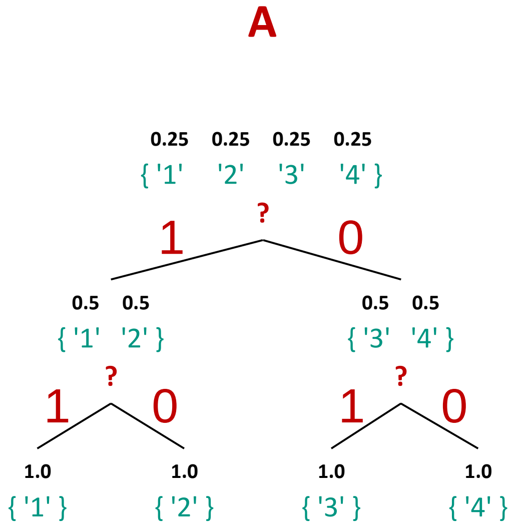
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

Basics

- Information

$$i(x) = -log_2(p(x))$$





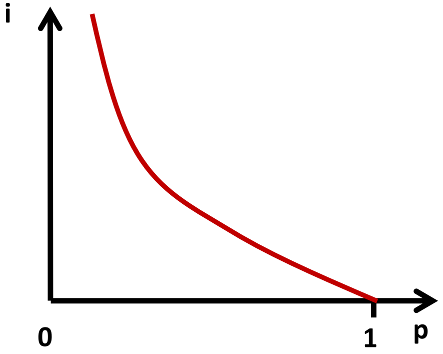
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

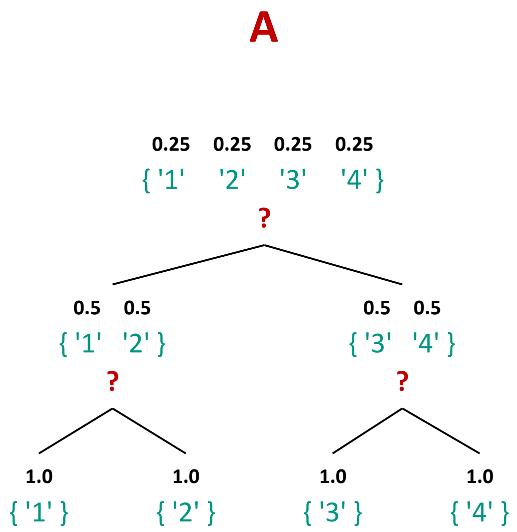
Any set of symbols can be expressed in a 1/0 alphabet …
… and can be interpreted as a sequence of answers on Yes/No questions
$H_2$0 = 1001000 110010 1001111

Basics

- Information

$$i(x) = -log_2\big(p(x)\big)$$





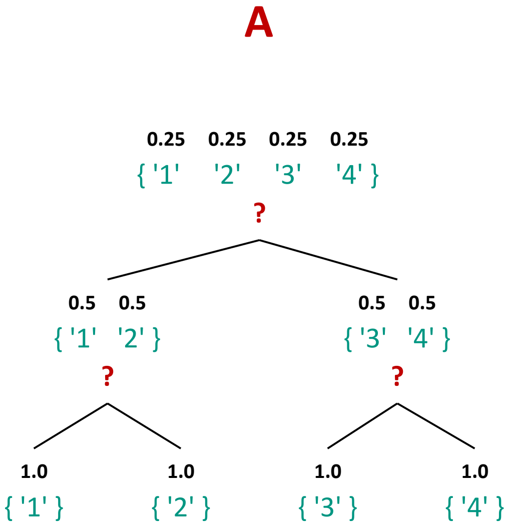
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

The smaller the probability of a signal, the more informative it is
Information is additive (probabilites are multiplicative)

Basics

- Entropy

$$i(x) = -log_2\big(p(x)\big)$$


A

0.25 0.25 0.25 0.25
{ '1' '2' '3' '4' }
?

0.5 0.5
{ '1' '2' }
?

0.5 0.5
{ '3' '4' }
?

1.0
{ '1' }

1.0
{ '2' }

1.0
{ '3' }

1.0
{ '4' }

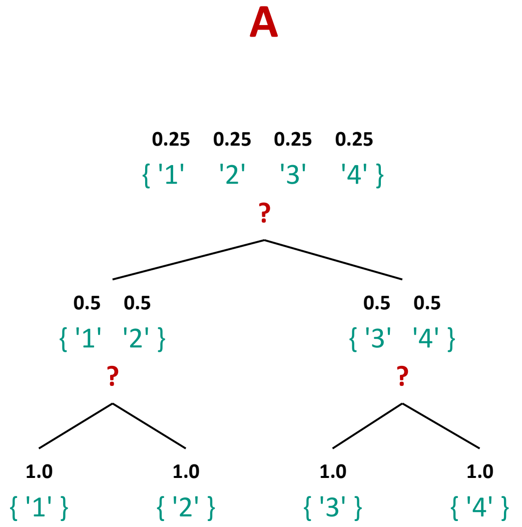
questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

Basics

- Entropy

$$i(x) = -log_2\left(p(x)\right)$$

$$H(X) = E(i(X)) = \sum_{j=1}^{n} i_j * p\left(x_j\right)$$

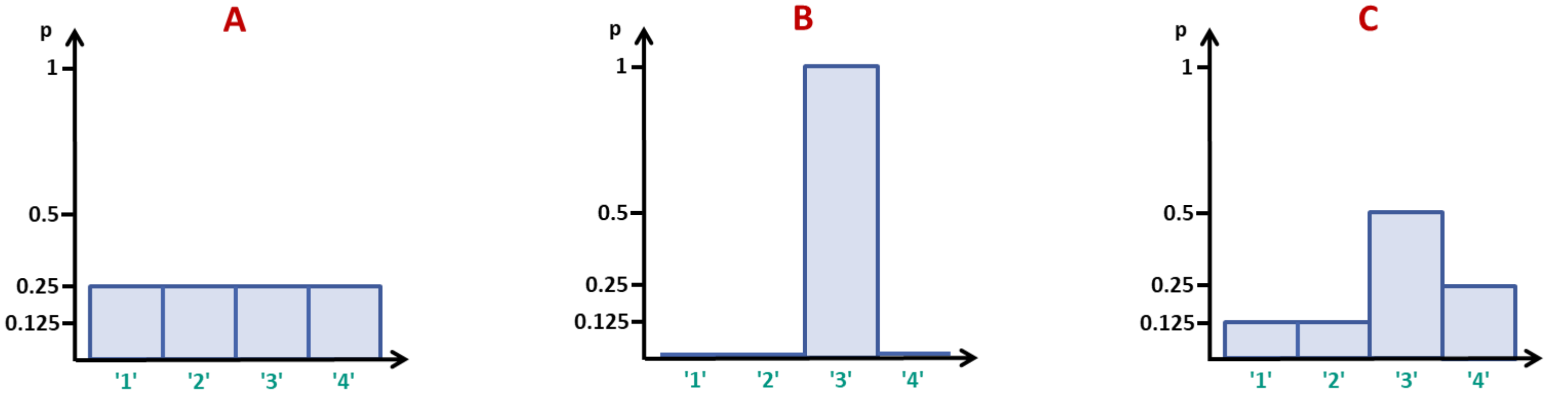


questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

Entropy = Average # of questions

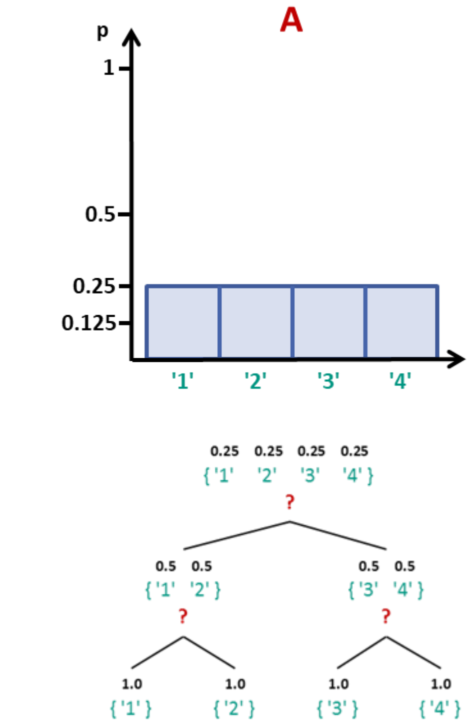
Basics

- Entropy

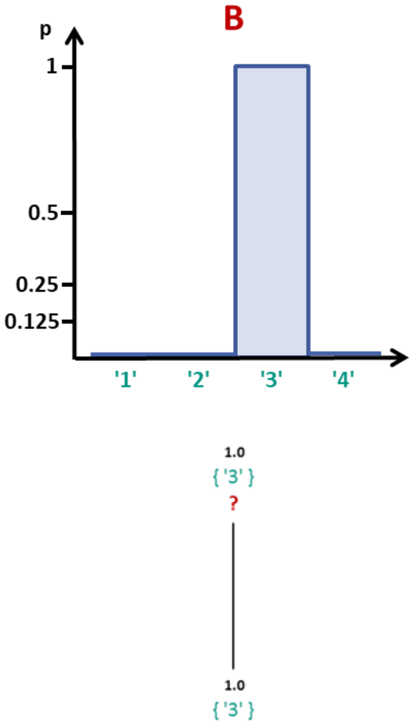


Basics

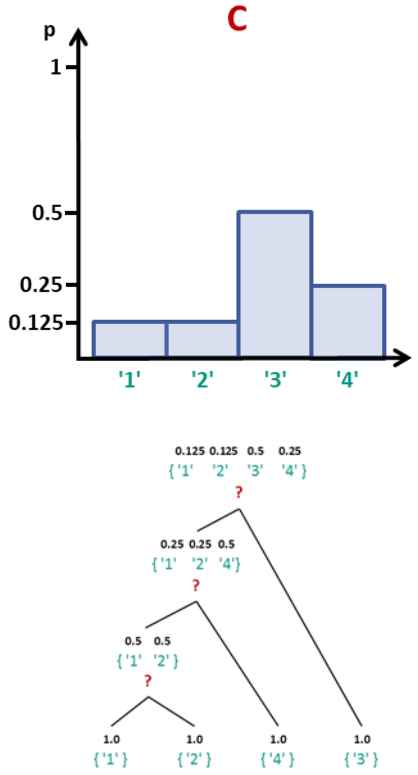
- Entropy



questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2



questions:	0
p:	1.0
i:	0



questions:	3	3	2	1
p:	0.125	0.125	0.25	0.5
i:	3	3	2	1

H(A) = 2 bit

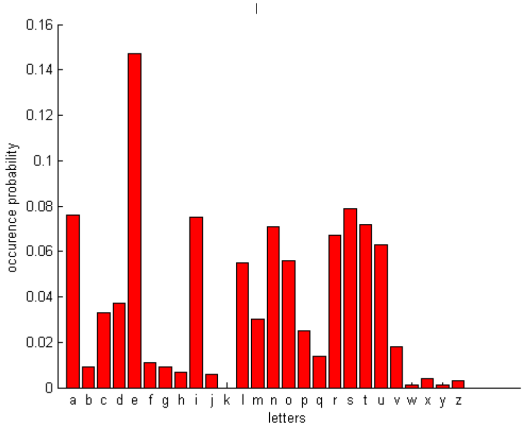
H(B) = 0 bit

H(C) = 1.75 bit

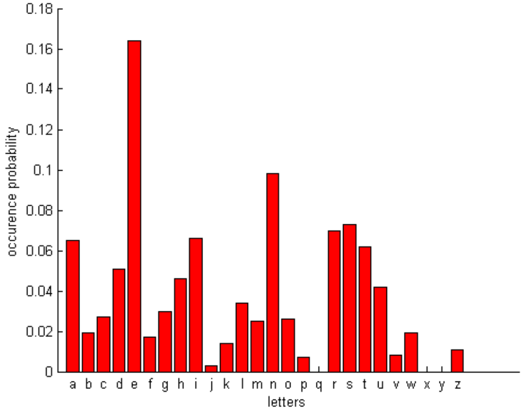
Basics

- How much information can we transmit with different alphabets?

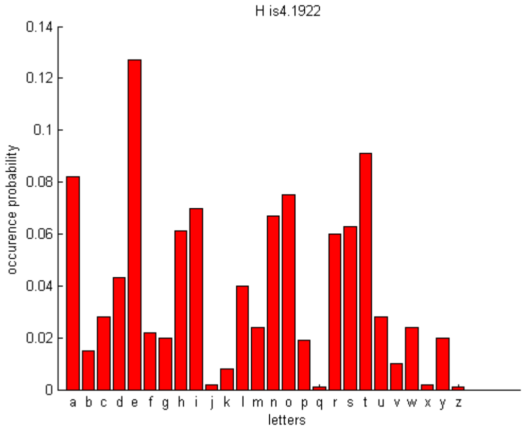
french



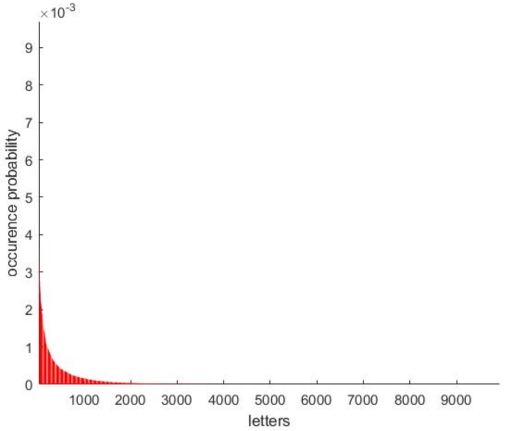
german



english



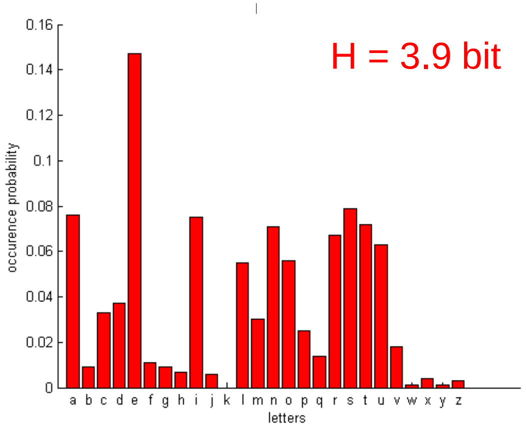
chinese



Basics

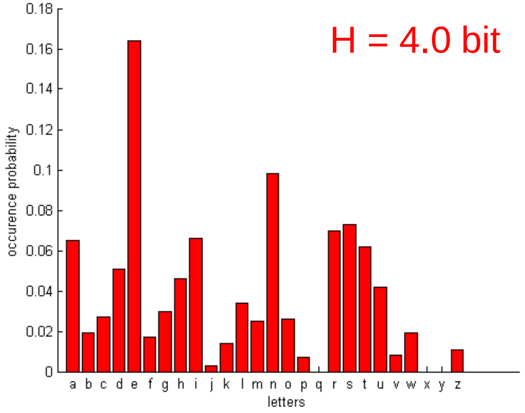
- How much information can we transmit with different alphabets?

french



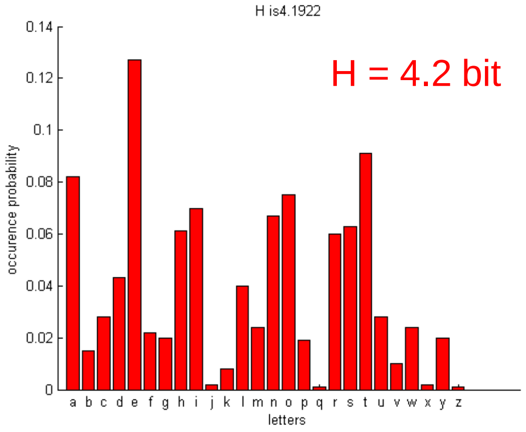
H = 3.9 bit

german



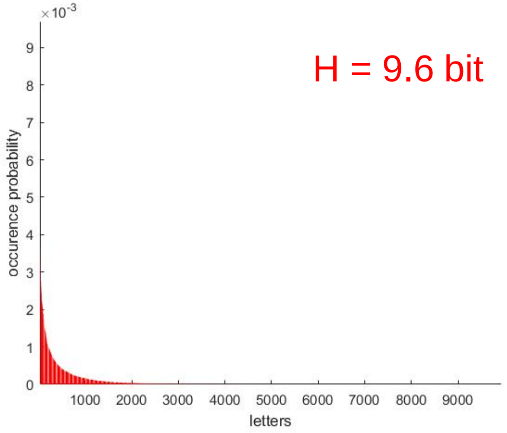
H = 4.0 bit

english



H = 4.2 bit

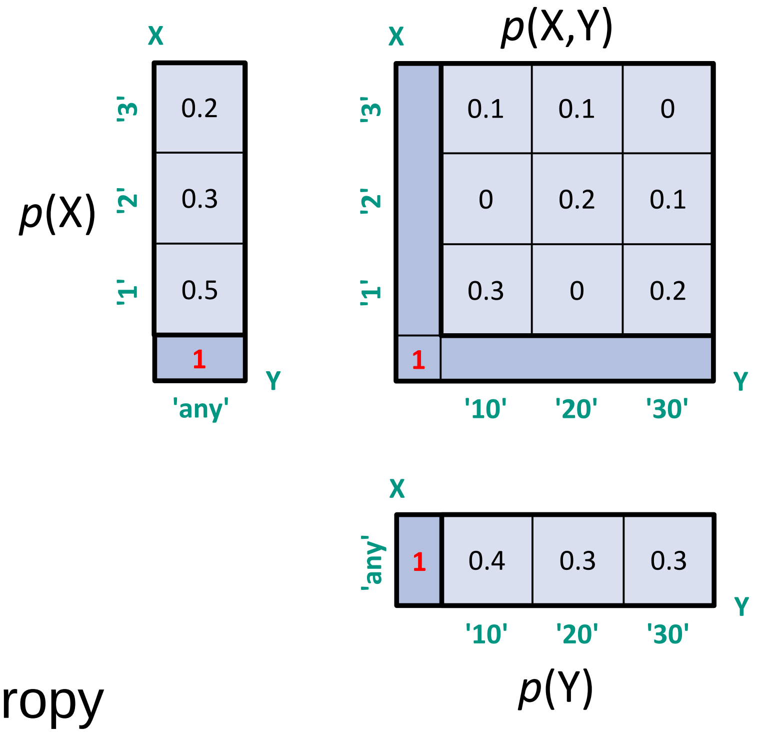
chinese



H = 9.6 bit

Basics

- What happens if the probability distribution is more than 1-d?



p(X,Y)

X \ Y	'10'	'20'	'30'
'3'	0.1	0.1	0
'2'	0	0.2	0.1
'1'	0.3	0	0.2
	1		

p(X)

X \ Y	'any'
'3'	0.2
'2'	0.3
'1'	0.5
	1

p(Y)

X \ Y		'10'	'20'	'30'
'any'	1	0.4	0.3	0.3

- Joint entropy is the entropy of a joint distribution
- Can be calculated for an any-dimensional distribution
- Can be interpreted as the average number of binary questions required to guess a paired signal (x,y)

- Joint Entropy

$$H(X,Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \cdot -log_2(p(x_i, y_j))$$

Basics

- What happens if we have prior knowledge of y when guessing x from (x,y)?

p(X)

X	
'3'	0.2
'2'	0.3
'1'	0.5
	1
Y	'any'

p(X,Y)

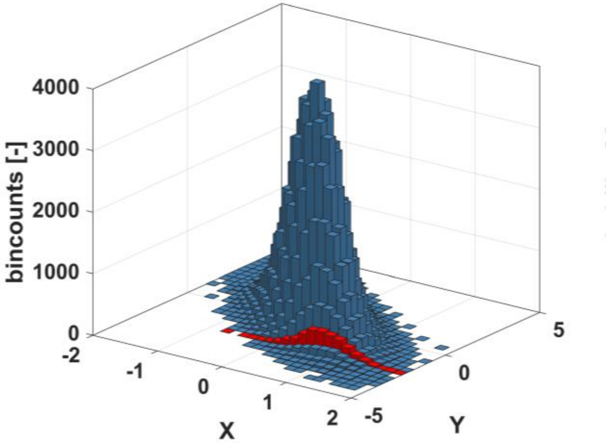
X				
'3'		0.1	0.1	0
'2'		0	0.2	0.1
'1'		0.3	0	0.2
	1			
Y		'10'	'20'	'30'

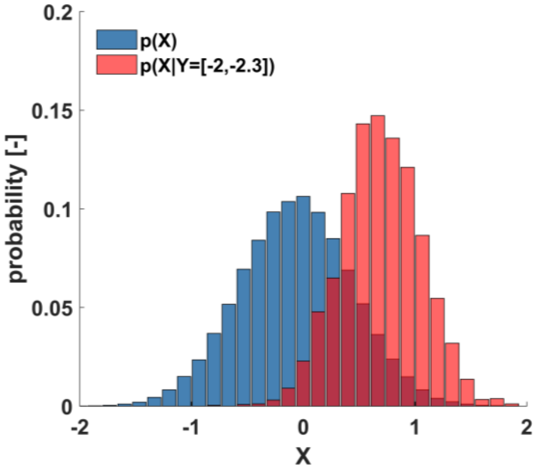
p(X|Y='10')

X	
'3'	0.25
'2'	0
'1'	0.75
	1
Y	'10'

X				
'any'	1	0.4	0.3	0.3
Y		'10'	'20'	'30'

p(Y)





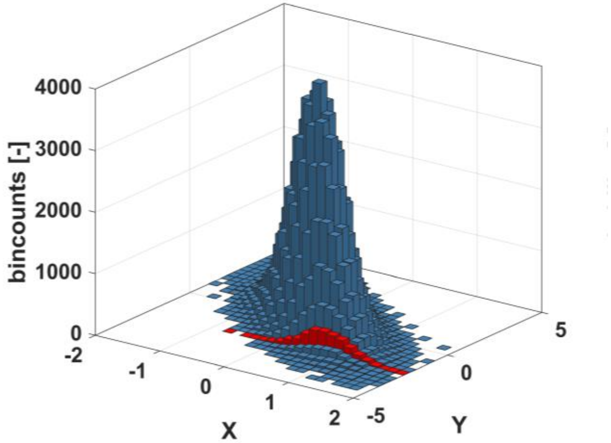
- Conditional Entropy

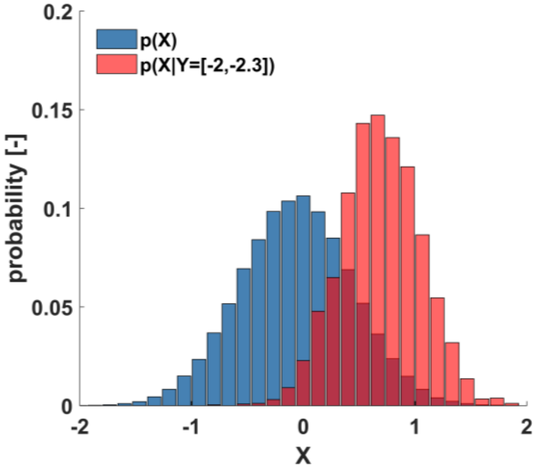
$$H(X|Y) = \sum_{j=1}^{n} p(y_j) \cdot H(X|Y = y_j)$$

Basics

- Information inequality
  - Compared to entropy $H(X)$ of the target, conditional entropy is the reduced uncertainty thanks to having advance knowledge of y when guessing x of a pair (x,y)
  - While for a particular y_j, $H(X|Y=y_j)$ can be $>$, $<$, or $= H(X)$, on average it holds

$$H(X) \geq H(X|Y)$$





Information can't hurt! (but there is a catch)

Basics

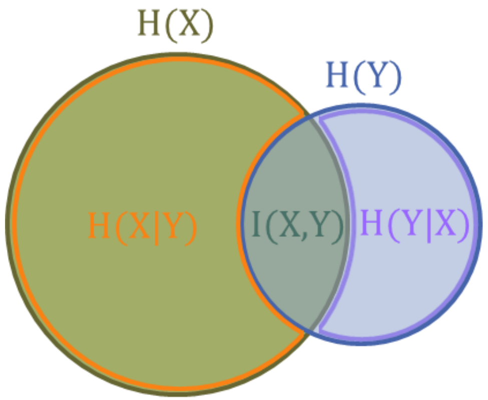
- The relation between Joint Entropy, Entropy, and Conditional Entropy

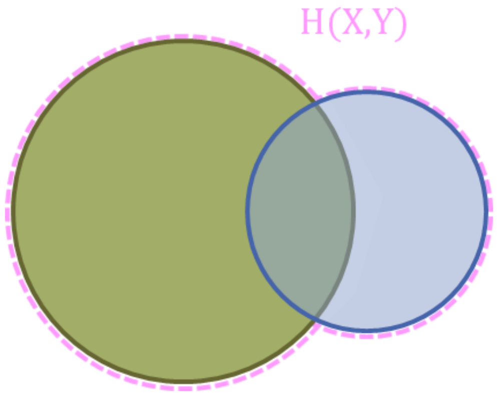
$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

- Mutual information

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

  - X knows in absolute terms as much about Y as Y knows about X!
  - In relative terms it may differ



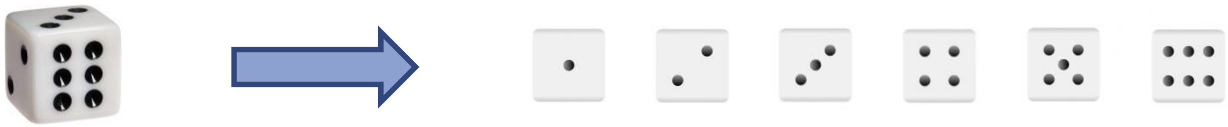


Basics

- Sources of uncertainty

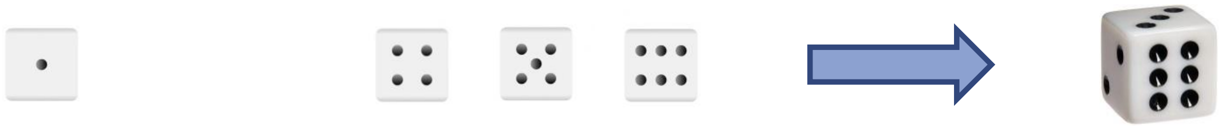
Basics

- Sources of uncertainty
  - Variability (system)



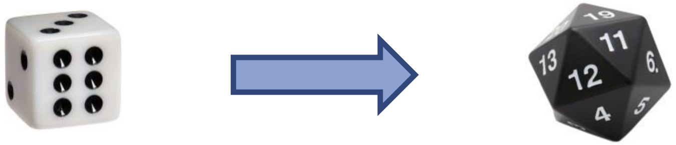
Basics

- Sources of uncertainty
  - Variability (system)
  - Representativity (sample size)



Basics

- Sources of uncertainty
  - Variability (system)
  - Representativity (sample size)
  - Consistency (application)



Basics

- Sources of uncertainty
  - Variability (system)
  - Representativity (sample size)
  - Consistency (application)
  - Compression (memory)

Basics

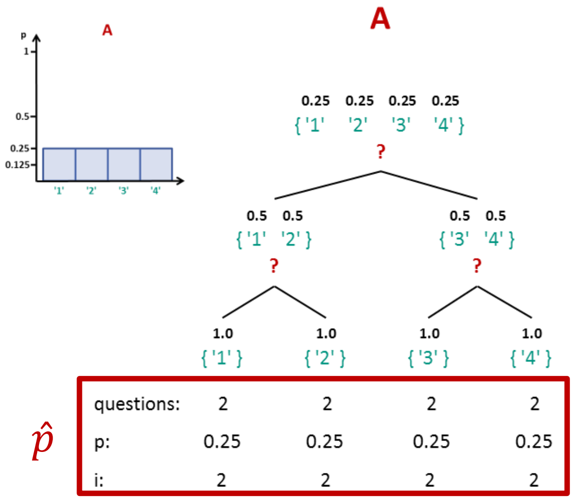
- Sources of uncertainty
  - Variability (system)
  - Representativity (sample size)
  - Consistency (application)
  - Compression (memory)

(Conditional) Entropy

?

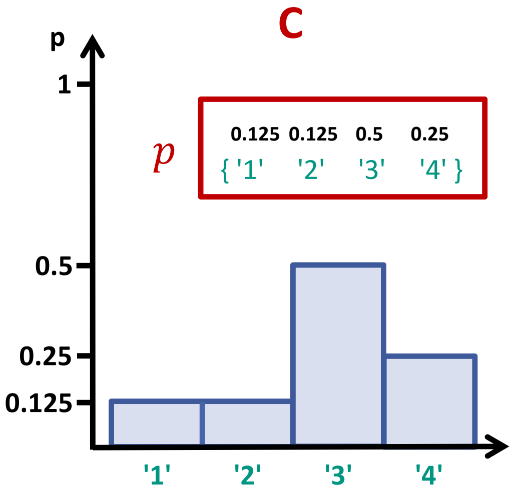
Basics

- Crossentropy



$\hat{p}$

questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2

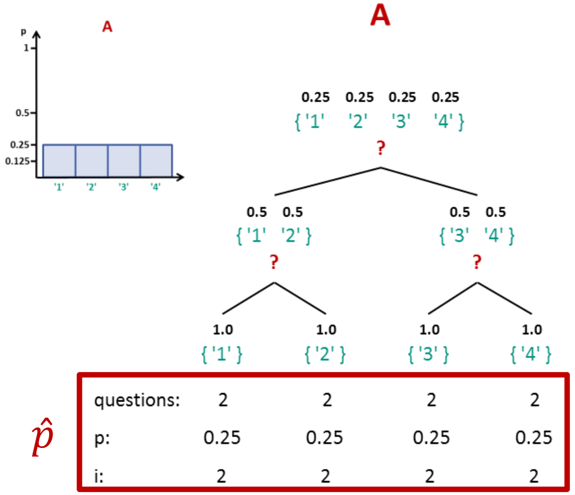


H(C) = 1.75 bit

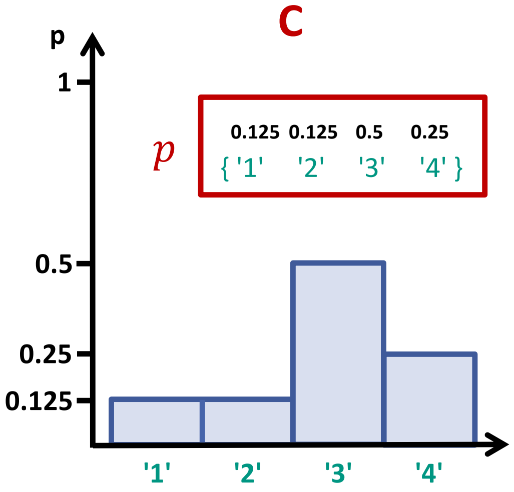
Basics

- Crossentropy

$$H_{cross}(X||\hat{X}) = \sum_{j=1}^{n} p(x_j) \cdot -log_2\left(\hat{p}(x_j)\right)$$



questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2



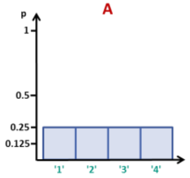
H(C) = 1.75 bit

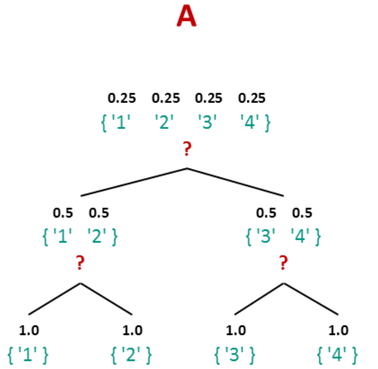
H(C||A) = 2 bit

Basics

- Crossentropy

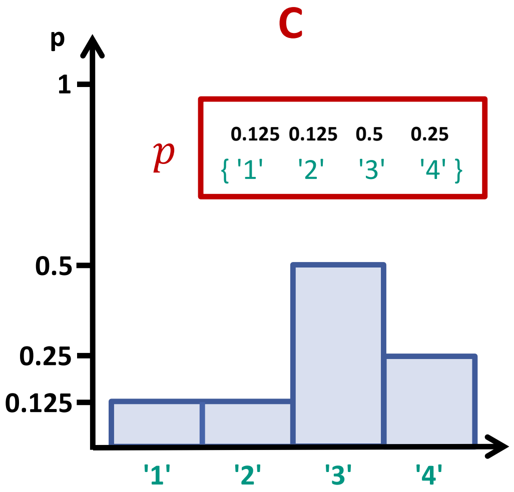
$$H_{cross}(X||\hat{X}) = \sum_{j=1}^{n} p(x_j) \cdot -log_2\left(\hat{p}(x_j)\right)$$





$\hat{p}$

questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2



H(C) = 1.75 bit

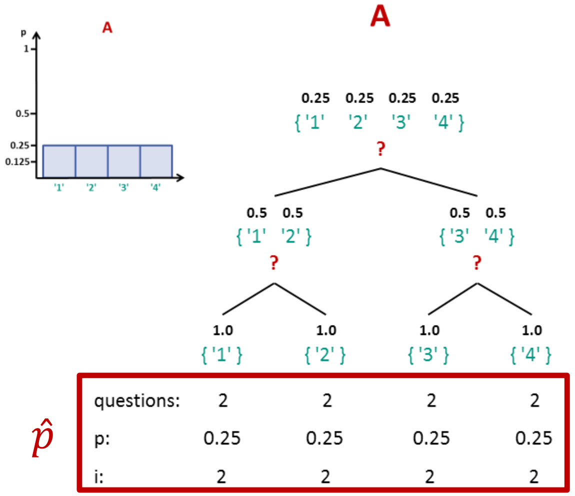
H(C||A) = 2 bit

Crossentropy measures total uncertainty from
Variability (data-uncertainty)
Representativity, Consistency, Compression (model-uncertainty)

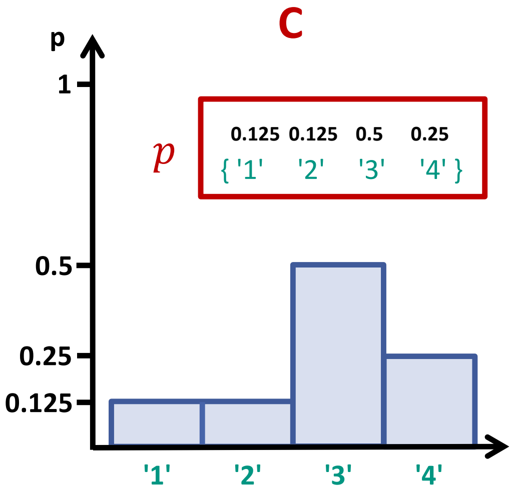
Basics

- Kullback-Leibler divergence

$$D_{KL}(X||\hat{X}) = H_{cross}(X||\hat{X}) - \mathrm{H(X)}$$



questions:	2	2	2	2
p:	0.25	0.25	0.25	0.25
i:	2	2	2	2



H(C) = 1.75 bit

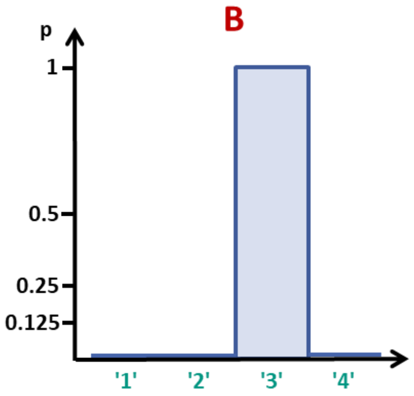
H(C||A) = 2 bit

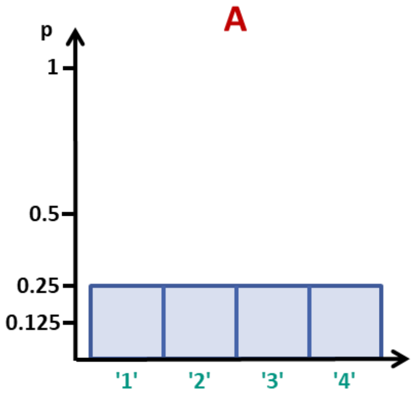
D_{KL}(C||A) = 0.25 bit

Kullback-Leibler divergence measures model-uncertainty

Limits

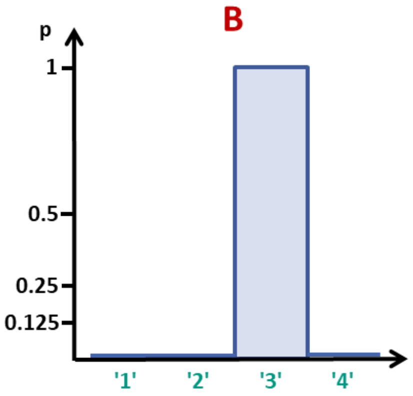
- Entropy limits

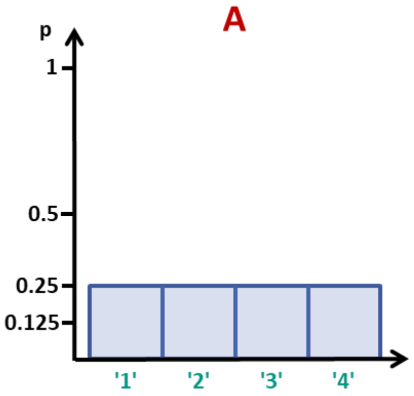




Limits

- Entropy limits



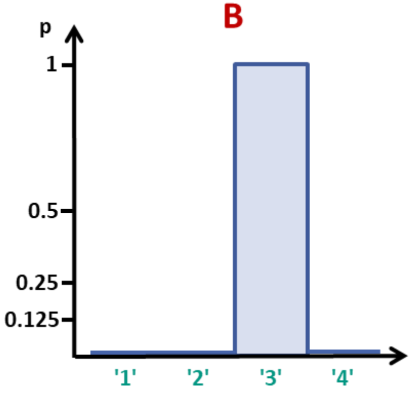


H(B) = 0 bit

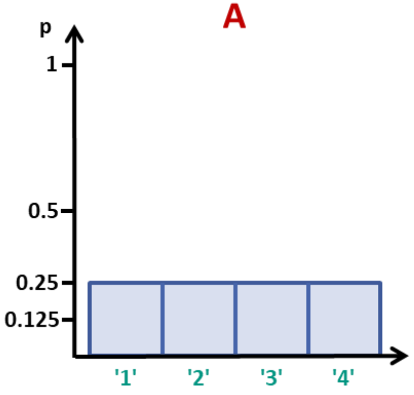
H(A) = 2 bit = $\log_2$(# of bins)

Limits

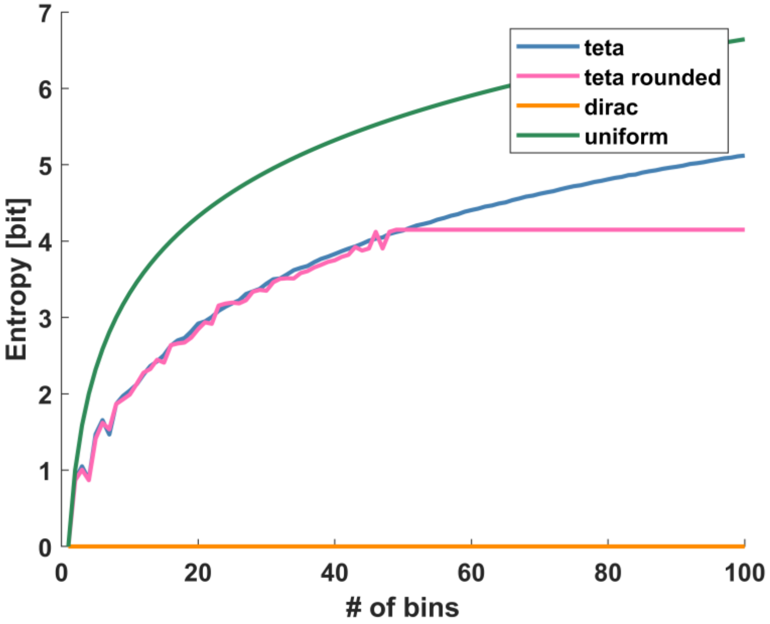
- Entropy limits



H(B) = 0 bit

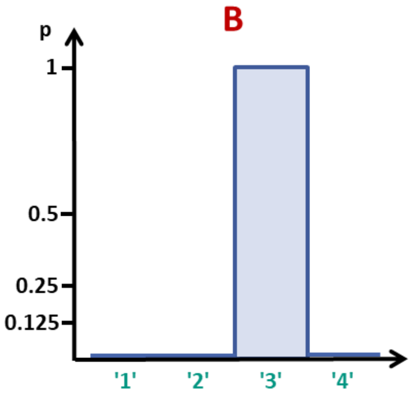


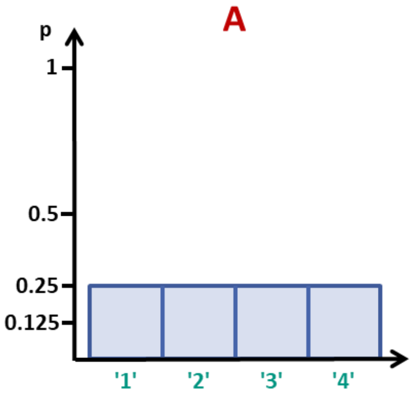
H(A) = 2 bit = $\log_2$(# of bins)

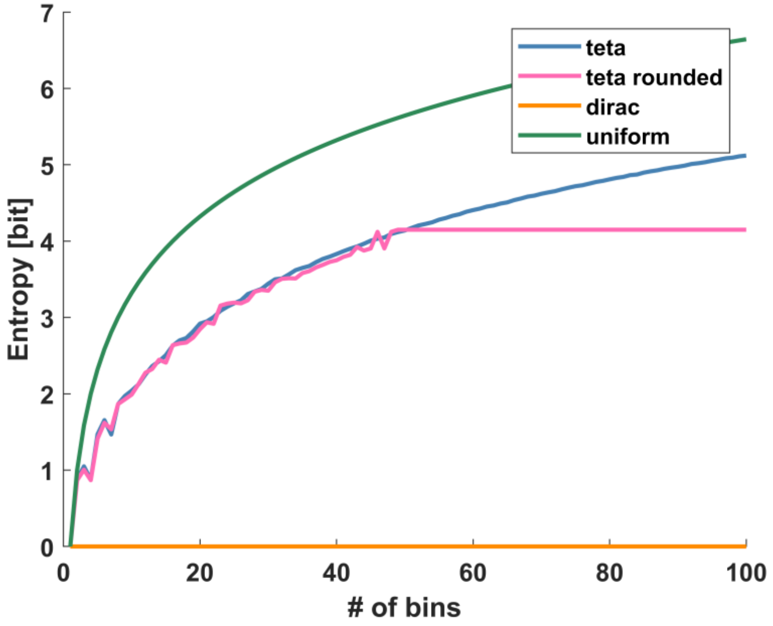


Limits

- Entropy limits







H(B) = 0 bit

H(A) = 2 bit = $\log_2$(# of bins)

- Maximum Entropy principle for inference
  - Jaynes (1957, 2003): "given all known constraints, all remaining unknowns should be represented by maximum entropy estimates"

Jaynes, E. T. (1957), Information Theory and Statistical Mechanics, Physical Review, 106(4), 620-630.
Jaynes, E. T. (2003), Probability Theory: The Logic of Science, Cambridge Univ. Press, Cambridge, UK.

Limits

- Koutsoyiannis (2014)
  - Uses elementary physical constraints and the Maximum Entropy Principle to derive ideal gas law, Clausius-Clapeyron relation, and others

Figure 2. Explanatory sketch indicating basic quantities involved in the equilibrium of the water vapour with liquid water, with zoom on a single molecule which "tries to hide itself" by maximizing the combined uncertainty related to its phase (being either gaseous or liquid with probabilities π_A and π_B, respectively), position and kinetic state.

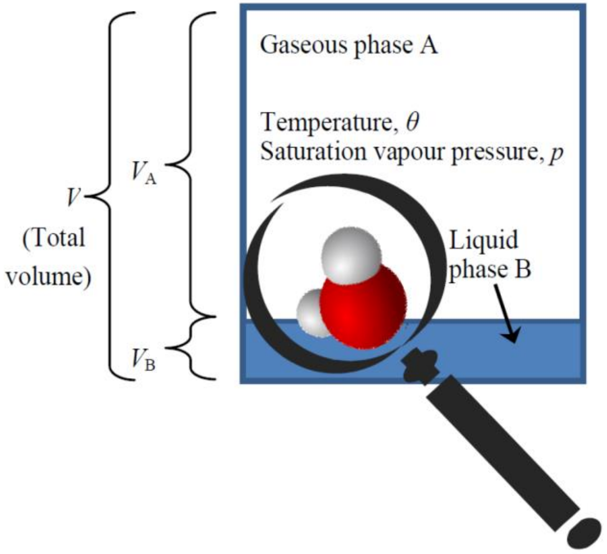
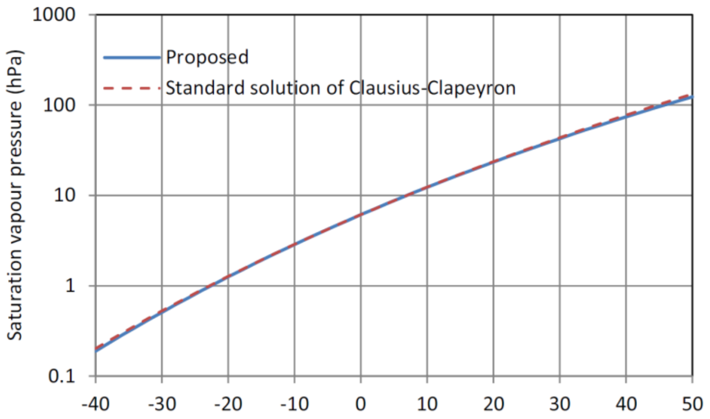


Figure 3. (**upper**) Comparison of saturation vapour pressure obtained by the proposed Equation (40) and by the standard but inconsistent Equation (41). (**lower**) Comparison of relative differences of the saturation vapour pressure obtained by the proposed Equation (40), as well as by Equation (41), with accurate measurement data of different origins, as indicated in the legend (for details on data see [34]).



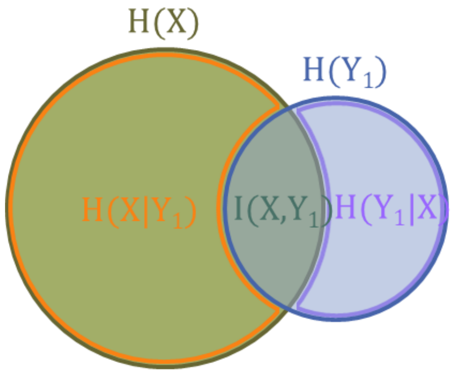
Koutsoyiannis, D. (2014), Entropy: From Thermodynamics to Hydrology, Entropy, 16(3), 1287.

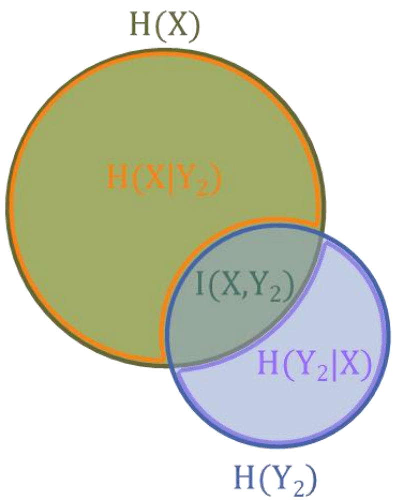
Independence, Redundancy, Synergy

- What happens to mutual information if more than one predictor exists?

- Remember mutual information

$I(X,Y) = H(X) - H(X|Y)$





Independence, Redundancy, Synergy

- What happens to mutual information if more than one predictor exists?

- Remember mutual information

  $I(X,Y) = H(X) - H(X|Y)$

- Independence

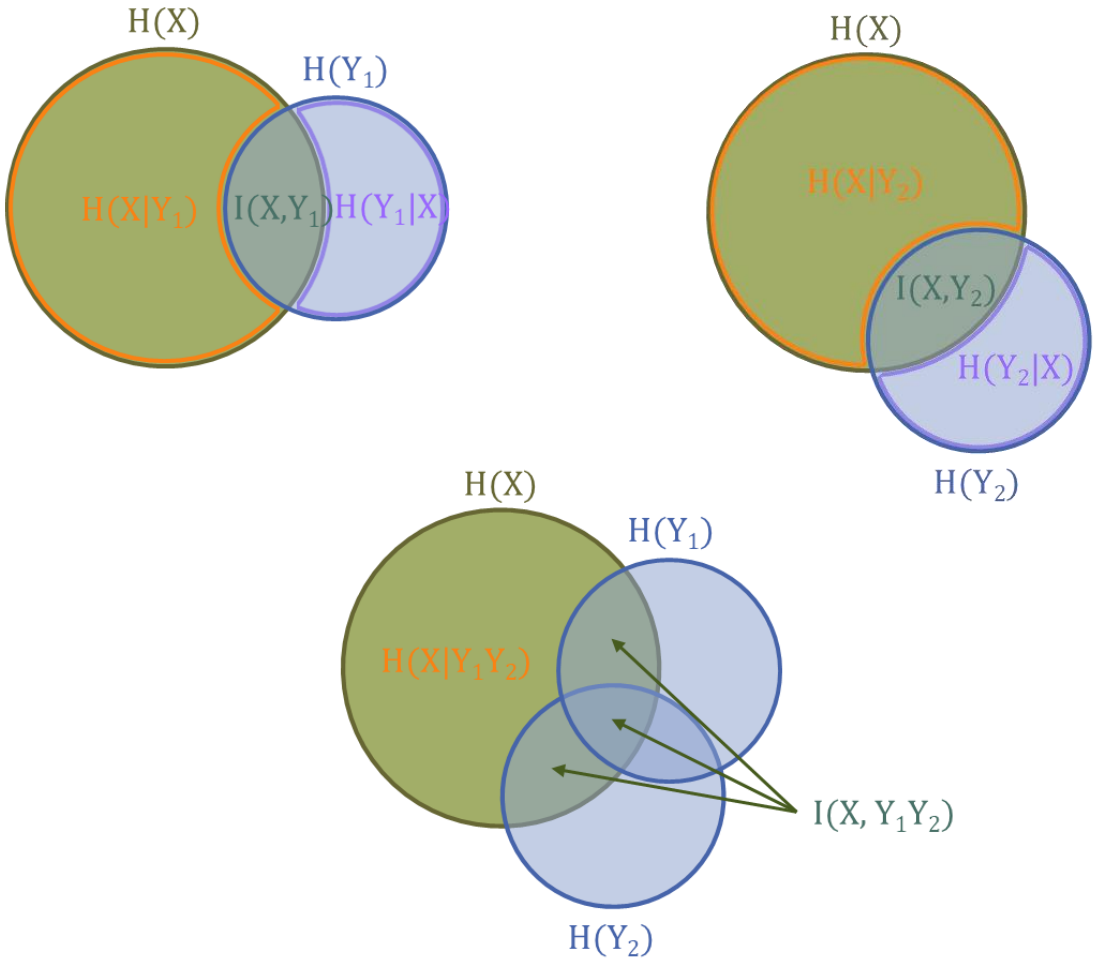
  $I(X,Y_1Y_2) = I(X,Y_1) + I(X,Y_2)$

- Redundancy

  $I(X,Y_1Y_2) < I(X,Y_1) + I(X,Y_2)$

- Synergy

  $I(X,Y_1Y_2) > I(X,Y_1) + I(X,Y_2)$



Independence, Redundancy, Synergy

Some data

Y1	Y2	X
1	1	1
1	2	1
2	1	2
2	2	2
1	1	3
2	1	3
1	2	4
2	2	4

$H(X) = 2$

$I(X,Y_1) = 0.5$

$I(X,Y_2) = 0.5$

$I(X,Y_1Y_2\ ) = 1$

$1 = 0.5 + 0.5$

→ independence

Independence, Redundancy, Synergy

Some data

Y1	Y2	X
1	1	1
1	2	1
2	1	2
2	2	2
1	1	3
2	1	3
1	2	4
2	2	4

$H(X) = 2$
$I(X,Y_1) = 0.5$
$I(X,Y_2) = 0.5$
$I(X,Y_1Y_2\,) = 1$
$1 = 0.5 + 0.5$
→ independence

CopyPaste

Y1	Y2	X
0	0	0
0	0	0
1	1	0
1	1	1

$H(X) = 0.81$
$I(X,Y_1) = 0.31$
$I(X,Y_2) = 0.31$
$I(X,Y_1Y_2\,) = 0.31$
$0.31 < 0.31 + 0.31$
→ redundancy

Independence, Redundancy, Synergy

Some data

Y1	Y2	X
1	1	1
1	2	1
2	1	2
2	2	2
1	1	3
2	1	3
1	2	4
2	2	4

$H(X) = 2$

$I(X,Y_1) = 0.5$

$I(X,Y_2) = 0.5$

$I(X,Y_1Y_2) = 1$

$1 = 0.5 + 0.5$

→ independence

CopyPaste

Y1	Y2	X
0	0	0
0	0	0
1	1	0
1	1	1

$H(X) = 0.81$

$I(X,Y_1) = 0.31$

$I(X,Y_2) = 0.31$

$I(X,Y_1Y_2) = 0.31$

$0.31 < 0.31 + 0.31$

→ redundancy

OR

Y1	Y2	X
0	0	0
0	1	1
1	0	1
1	1	1

$H(X) = 0.81$

$I(X,Y_1) = 0.31$

$I(X,Y_2) = 0.31$

$I(X,Y_1Y_2) = 0.81$

$0.81 > 0.31 + 0.31$

→ synergy

Independence, Redundancy, Synergy

Some data

Y1	Y2	X
1	1	1
1	2	1
2	1	2
2	2	2
1	1	3
2	1	3
1	2	4
2	2	4

$H(X) = 2$
$I(X,Y_1) = 0.5$
$I(X,Y_2) = 0.5$
$I(X,Y_1Y_2) = 1$
$1 = 0.5 + 0.5$
→ independence

CopyPaste

Y1	Y2	X
0	0	0
0	0	0
1	1	0
1	1	1

$H(X) = 0.81$
$I(X,Y_1) = 0.31$
$I(X,Y_2) = 0.31$
$I(X,Y_1Y_2) = 0.31$
$0.31 < 0.31 + 0.31$
→ redundancy

OR

Y1	Y2	X
0	0	0
0	1	1
1	0	1
1	1	1

$H(X) = 0.81$
$I(X,Y_1) = 0.31$
$I(X,Y_2) = 0.31$
$I(X,Y_1Y_2) = 0.81$
$0.81 > 0.31 + 0.31$
→ synergy

XOR

Y1	Y2	X
0	0	0
0	1	1
1	0	1
1	1	0

$H(X) = 1$
$I(X,Y_1) = 0$
$I(X,Y_2) = 0$
$I(X,Y_1Y_2) = 1$
$1 > 0 + 0$
→ synergy

Independence, Redundancy, Synergy

Some data

Y1	Y2	X
1	1	1
1	2	1
2	1	2
2	2	2
1	1	3
2	1	3
1	2	4
2	2	4

$H(X) = 2$
$I(X,Y_1) = 0.5$
$I(X,Y_2) = 0.5$
$I(X,Y_1Y_2\,) = 1$
$1 = 0.5 + 0.5$
→ independence

CopyPaste

Y1	Y2	X
0	0	0
0	0	0
1	1	0
1	1	1

$H(X) = 0.81$
$I(X,Y_1) = 0.31$
$I(X,Y_2) = 0.31$
$I(X,Y_1Y_2\,) = 0.31$
$0.31 < 0.31 + 0.31$
→ redundancy

OR

Y1	Y2	X
0	0	0
0	1	1
1	0	1
1	1	1

$H(X) = 0.81$
$I(X,Y_1) = 0.31$
$I(X,Y_2) = 0.31$
$I(X,Y_1Y_2\,) = 0.81$
$0.81 > 0.31 + 0.31$
→ synergy

XOR

Y1	Y2	X
0	0	0
0	1	1
1	0	1
1	1	0

$H(X) = 1$
$I(X,Y_1) = 0$
$I(X,Y_2) = 0$
$I(X,Y_1Y_2\,) = 1$
$1 > 0 + 0$
→ synergy

I, R, S can jointly occur in the same data set and compensate

Conservation, Dissipation, Innovation

- Data processing inequality
  - "…no processing of Y, deterministic or random, can increase the information that Y contains about X." (Cover and Thomas, 2006)

$$H(X|Y) \leq H(X|f(Y))$$

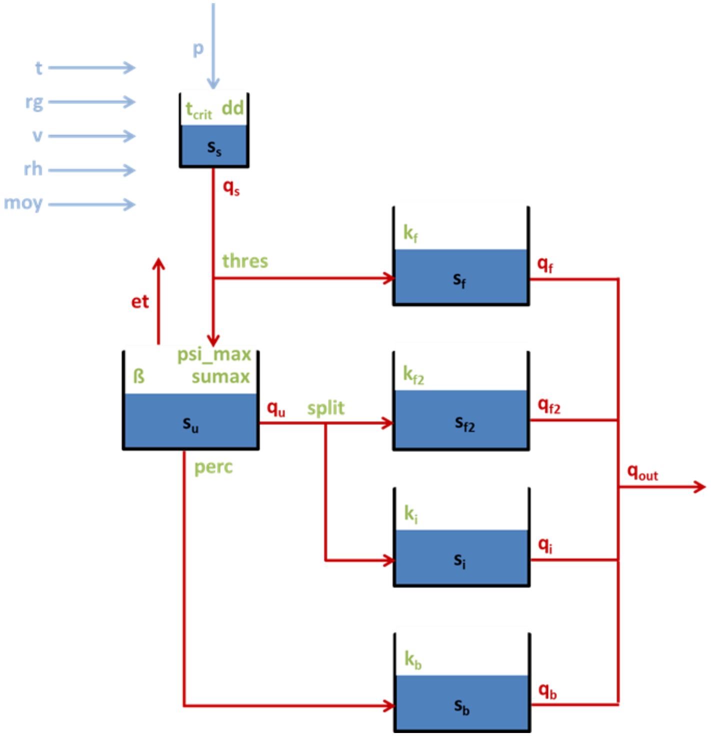
Cover, T., and Thomas, J. A. (2006): Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, 2006

Conservation, Dissipation, Innovation

- Data processing inequality
  - "…no processing of Y, deterministic or random, can increase the information that Y contains about X." (Cover and Thomas, 2006)

$$H(X|Y) \leq H(X|f(Y))$$

$$\begin{tikzcd}
q_{out} = f(q_i) & q_i = f(q_u) \arrow[l, bend left] & q_u = f(q_s) \arrow[l, bend left] & q_s = f(p) \arrow[l, bend left]
\end{tikzcd}$$



Cover, T., and Thomas, J. A. (2006): Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, 2006

Conservation, Dissipation, Innovation

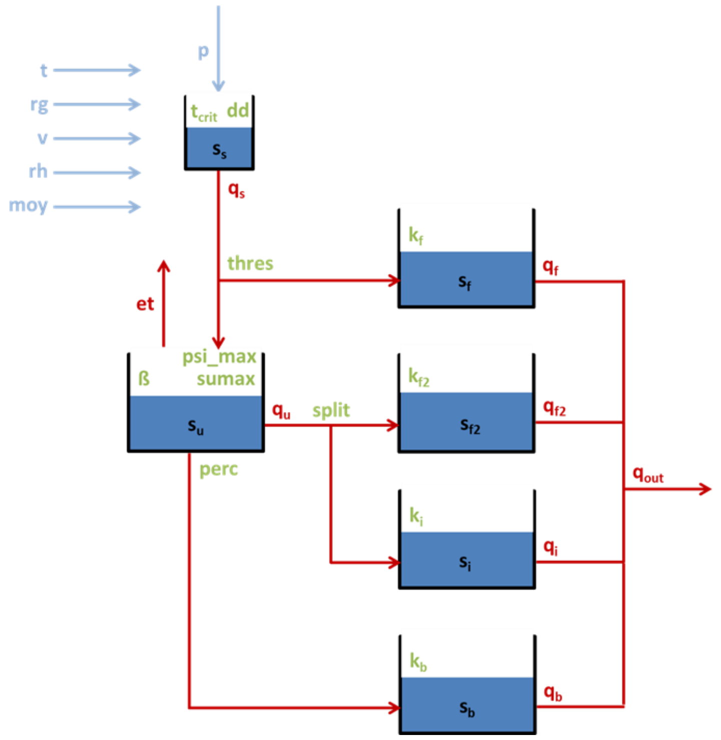
- Data processing inequality
  - "…no processing of Y, deterministic or random, can increase the information that Y contains about X." (Cover and Thomas, 2006)

$$H(X|Y) \leq H(X|f(Y))$$

$$q_{out} = f(q_i) \qquad q_i = f(q_u) \qquad q_u = f(q_s) \qquad q_s = f(p)$$

Why building models as series of data transformations, if no information is gained?

$$q_{out} = f(q_i) \qquad q_i = f(q_u) \qquad q_u = f(q_s) \qquad q_s = f(p)$$



Cover, T., and Thomas, J. A. (2006): Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, 2006

Conservation, Dissipation, Innovation

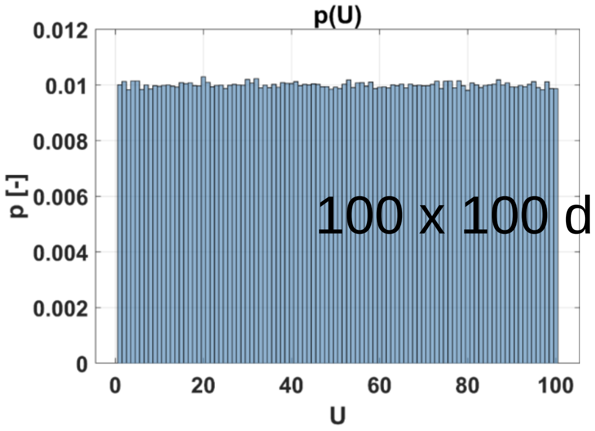
- Variability and Innovation

Conservation, Dissipation, Innovation

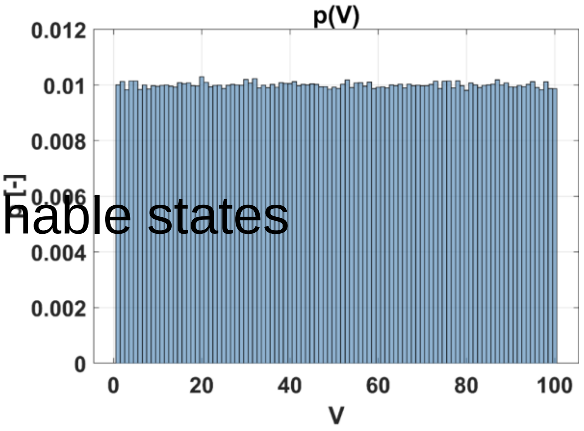
- Variability and Innovation

$$X = U + V$$

  - H(U,V) = 13.28 bit



100 x 100 distinguishable states

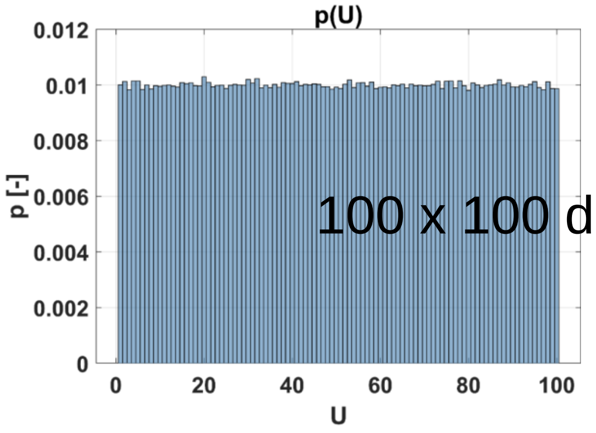


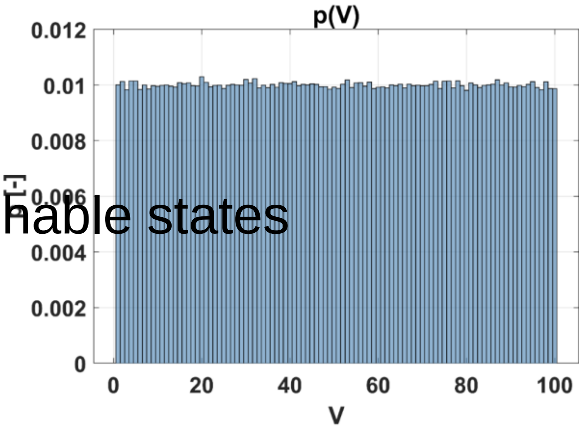
Conservation, Dissipation, Innovation

- Variability and Innovation

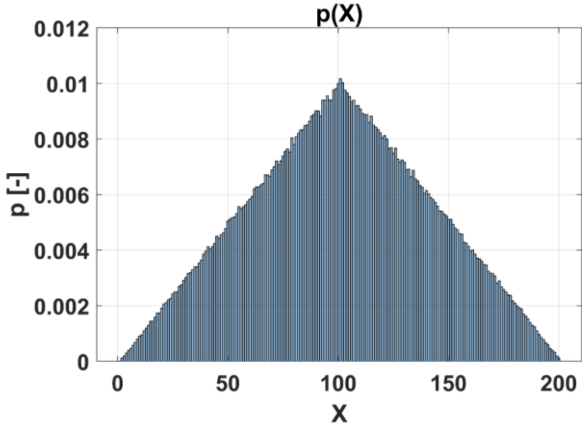
$$X = U + V$$

  - H(U,V) = 13.28 bit
  - H(X|U,V) = 0 bit





100 x 100 distinguishable states

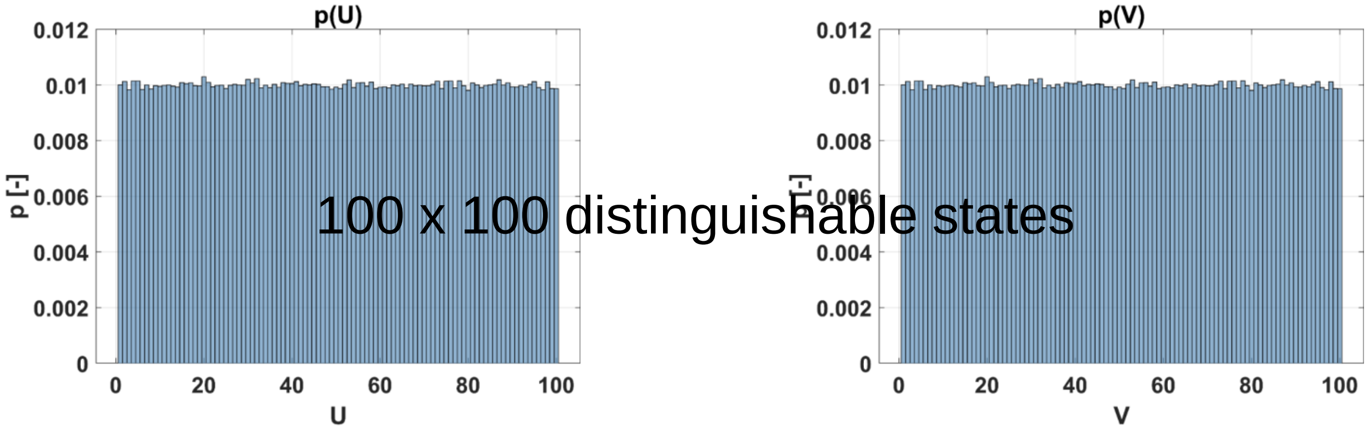


Conservation, Dissipation, Innovation

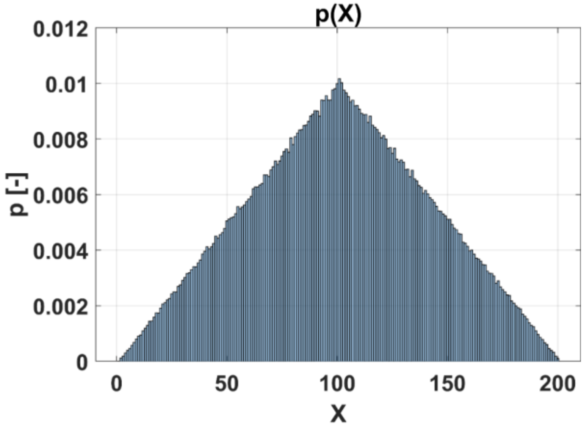
- Variability and Innovation

$$X = U + V$$

  - H(U,V) = 13.28 bit
  - H(X|U,V) = 0 bit



100 x 100 distinguishable states



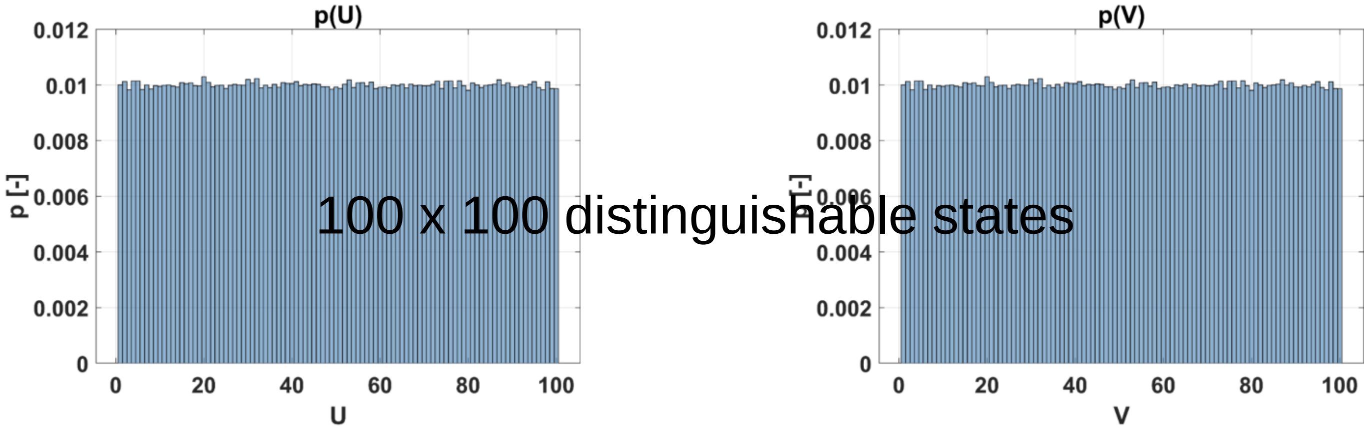
200 distinguishable states

Conservation, Dissipation, Innovation

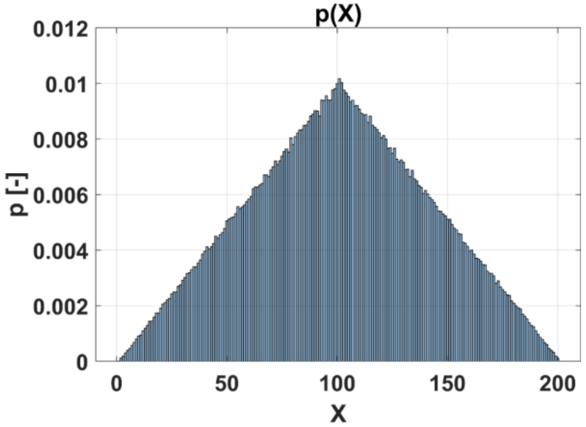
- Variability and Innovation

  $$X = U + V$$

  - H(U,V) = 13.28 bit
  - H(X|U,V) = 0 bit
  - H(X) = 7.36 bit < 13.28 bit!



100 x 100 distinguishable states



200 distinguishable states

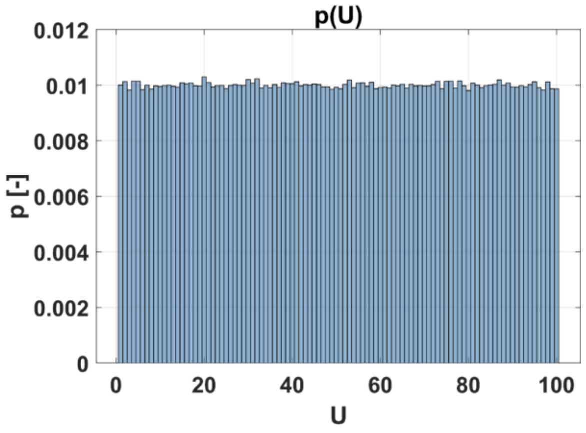
Conservation, Dissipation, Innovation

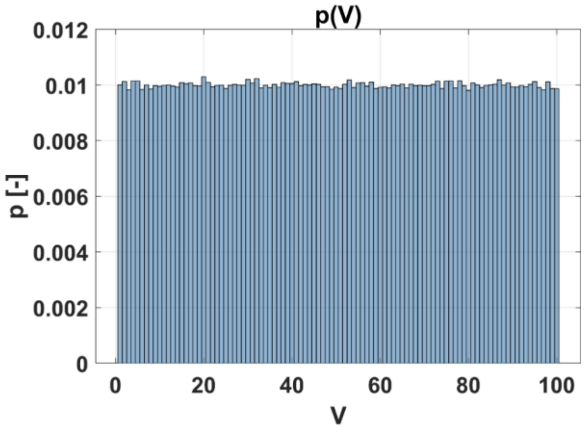
- Variability and Innovation

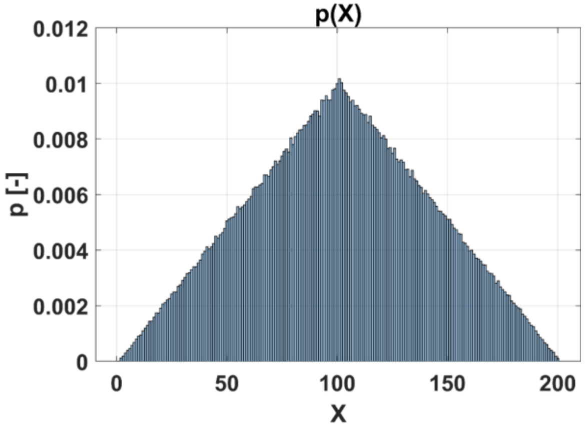
$$X = U + V$$

  - H(U,V) = 13.28 bit
  - H(X|U,V) = 0 bit
  - H(X) = 7.36 bit < 13.28 bit!

$$H(X = f(U,V)) \leq H(U,V)$$







Are things getting ever more boring?

How then does a random-number generator work?

Conservation, Dissipation, Innovation

Why building models if no information is gained?

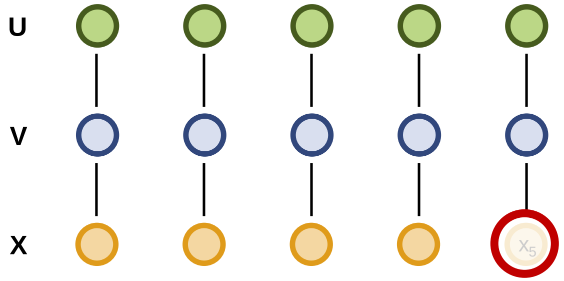
Are things getting ever more boring?

How then does a random-number generator work?

Conservation, Dissipation, Innovation

- Independent events

$X = f(U, V)$ ☹



Why building models if no information is gained?

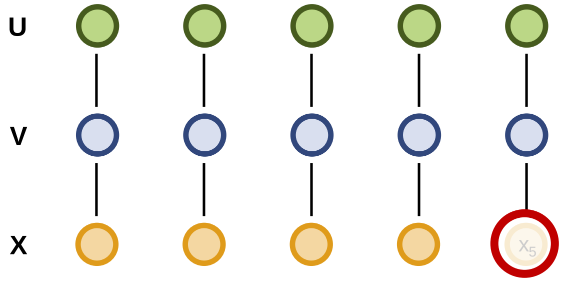
Are things getting ever more boring?

How then does a random-number generator work?

Conservation, Dissipation, Innovation

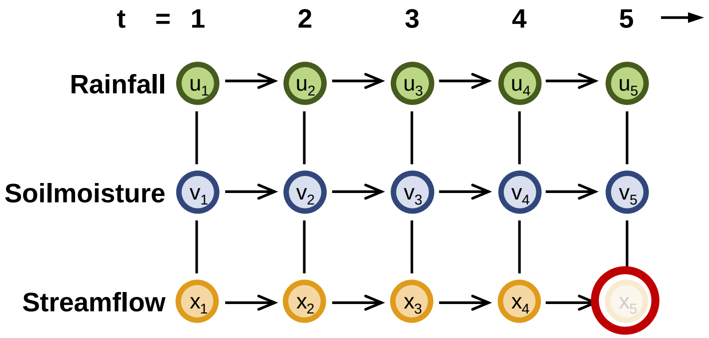
- Independent events

$$X = f(U,V)$$ ☹



- Order & memory does the trick!

$$X_t = f(U,V,X_{t-1})$$ ☺

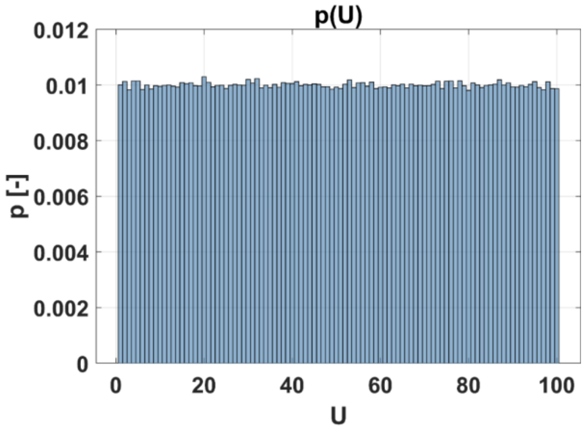

t = 1 2 3 4 5
Rainfall
u_1 u_2 u_3 u_4 u_5
Soilmoisture
v_1 v_2 v_3 v_4 v_5
Streamflow
x_1 x_2 x_3 x_4 x_5

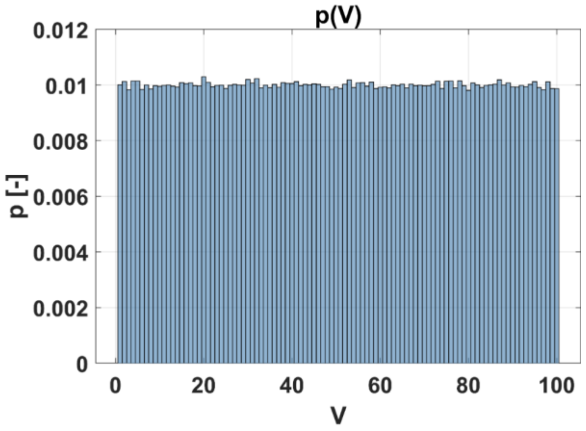
Conservation, Dissipation, Innovation

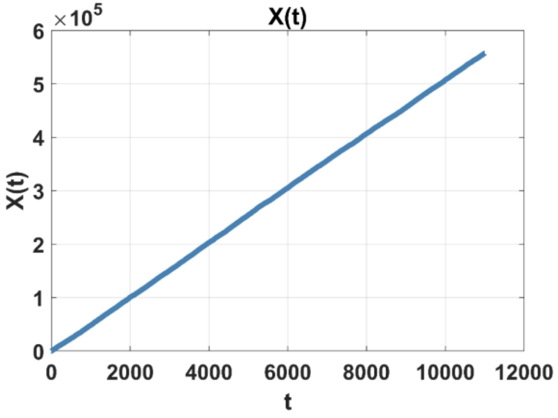
- Order & Memory!

$$X_t = U + V + X_{t-1}$$

  - H(U,V) = 13.28 bit







Conservation, Dissipation, Innovation

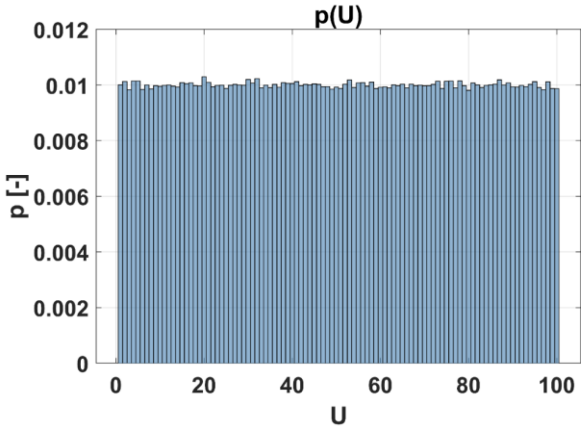
- Order & Memory!

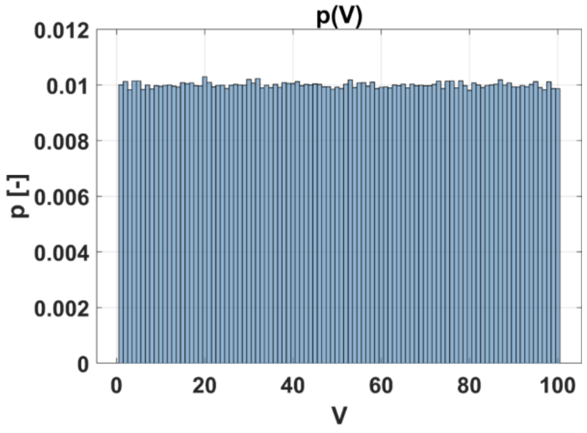
  $$X_t = U + V + X_{t-1}$$

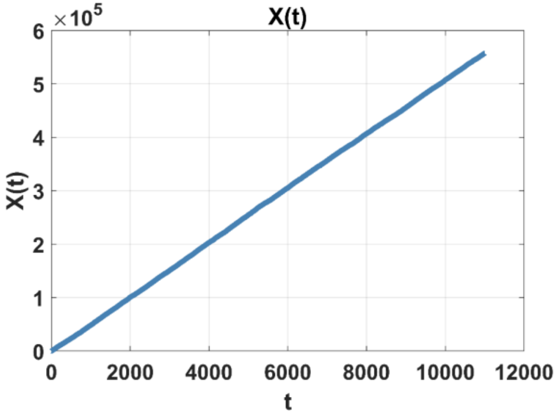
  - H(U,V) = 13.28 bit

    after 11000 timesteps
  - H(X) = 13.42 bit **>** 13.28 bit!

$$H(X_t = f(U,V,X_{t-1})) > H(U,V)$$ is possible!

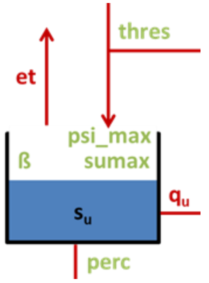




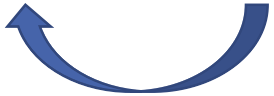


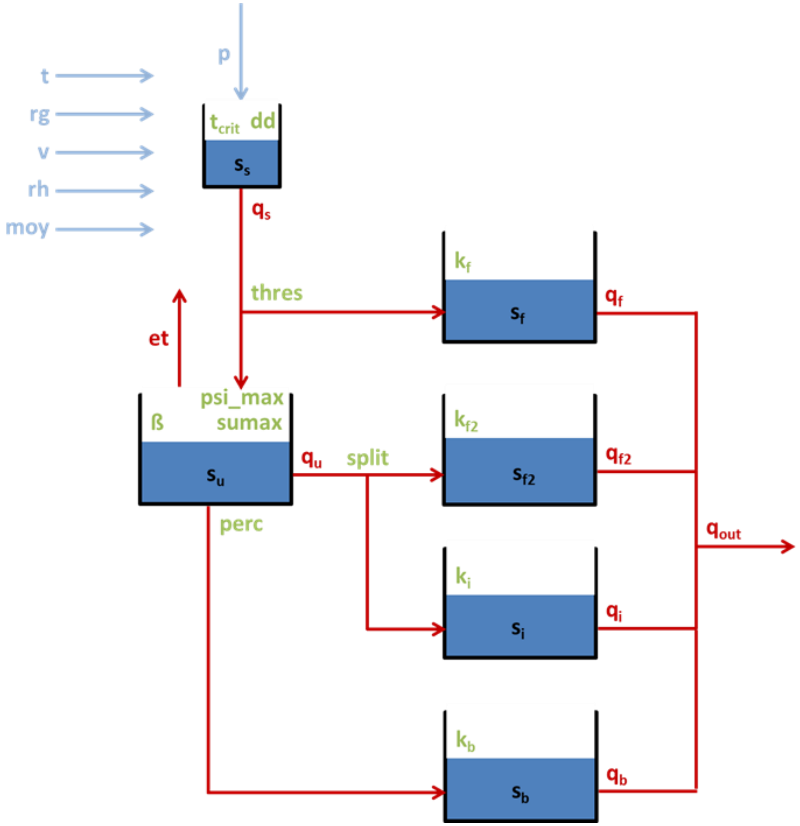
Conservation, Dissipation, Innovation

- Equation-based models
  - Recursion/Memory terms!



$$s_{u,t} = s_{u,t-1} + q_{s,t} - et - q_u - q_{perc}$$

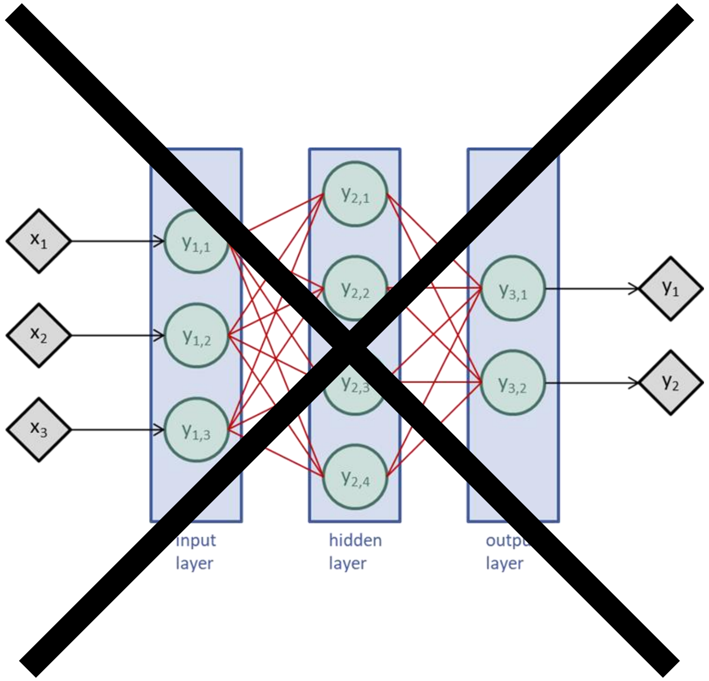




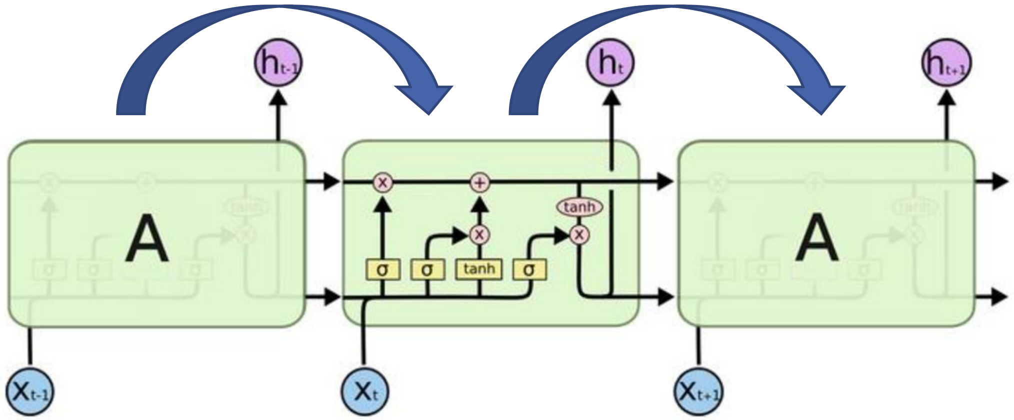
Conservation, Dissipation, Innovation

- Data-based models

Neural Network



Long Short-Term Memory (LSTM) Network



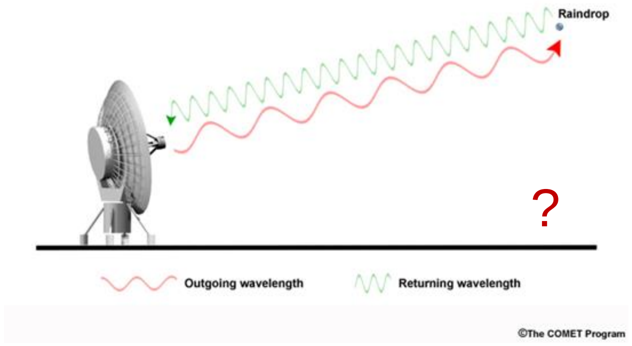
Conservation, Dissipation, Innovation

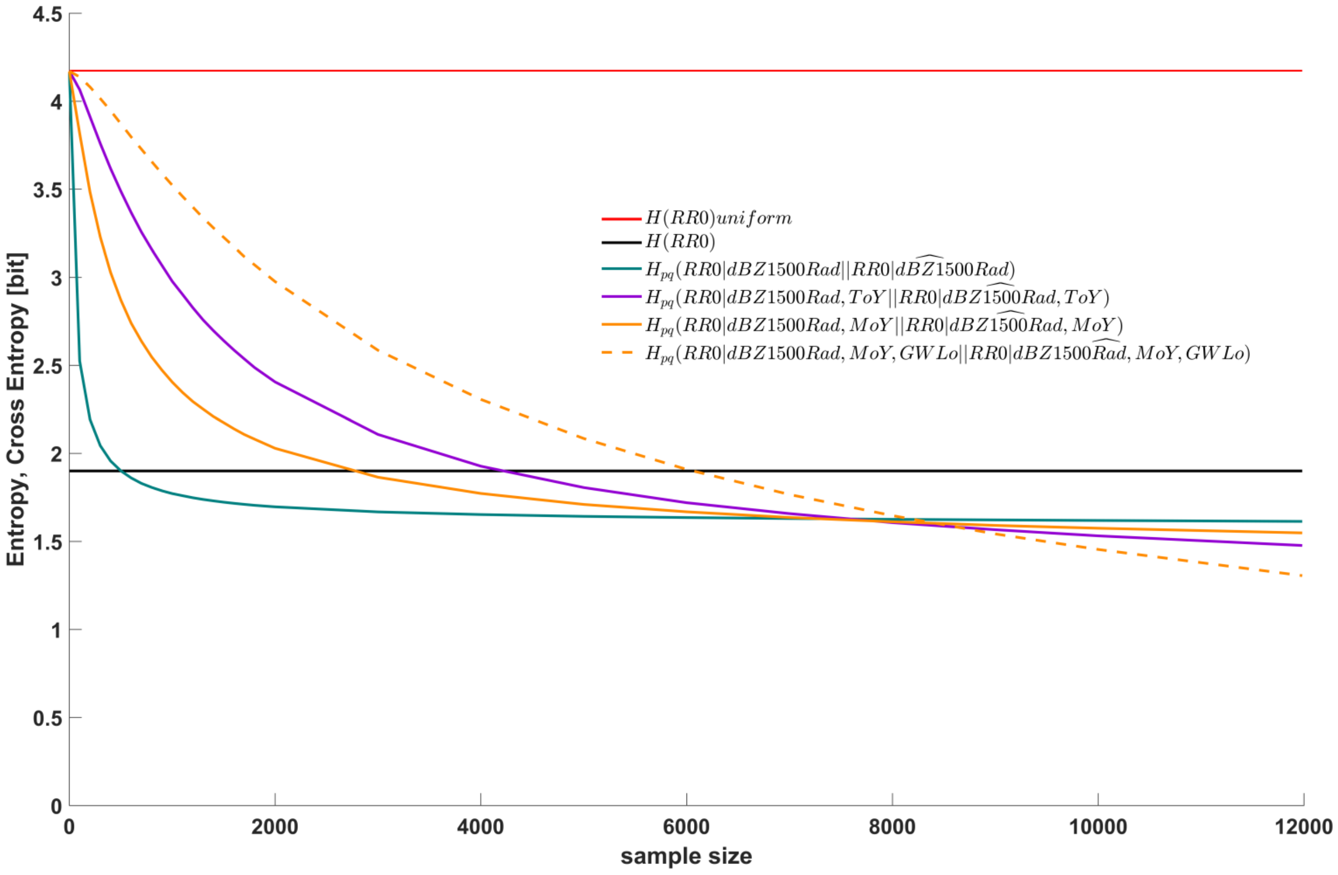
- Random-number generator
  - Linear congruential generator

$$x_t = (a \cdot x_{t-1} + b) \, mod \, m$$

Applications

- Neuper and Ehret (2019)
  - Quantify the information effect of limited sample size and number of predictors

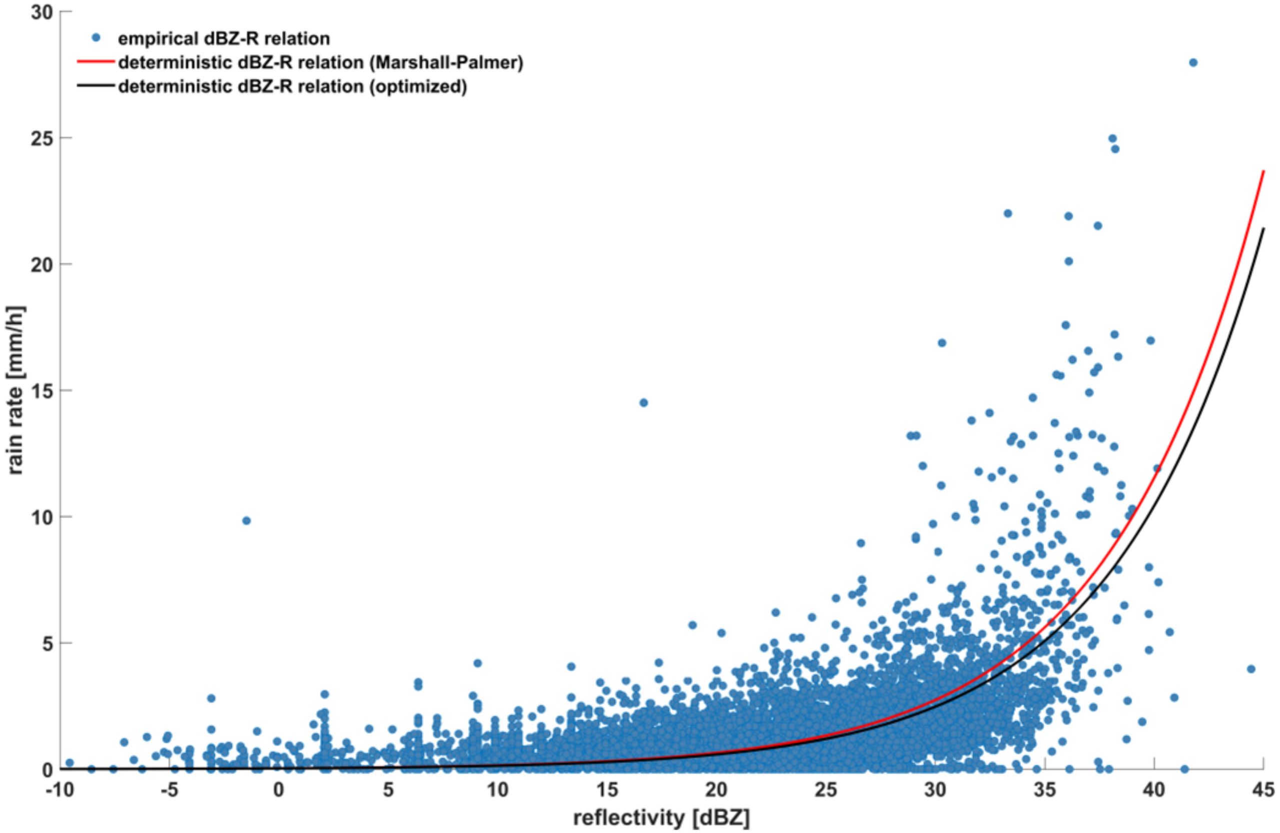




Neuper, M., and U. Ehret (2019): Quantitative precipitation estimation with weather radar using a data- and information-based approach, Hydrol. Earth Syst. Sci., 23(9), 3711-3733.

Applications

- Neuper and Ehret (2019)
  - Quantify the information effect of deterministic compression
    - H(RR) = 1.90 bit
    - H(RR|dBZ) = 1.61 bit
    - H_{cross}(RR|dBZ||Marshall-Palmer) = **5.04 bit** (**!**)

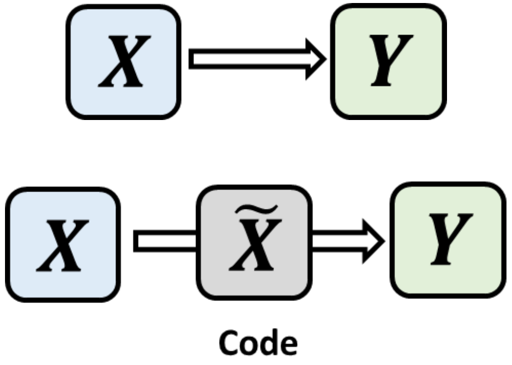


Neuper, M., and U. Ehret (2019): Quantitative precipitation estimation with weather radar using a data- and information-based approach, Hydrol. Earth Syst. Sci., 23(9), 3711-3733.

Applications

- Tishby et al. (2000)
  - Propose an information-based mechanism for feature extraction in ML

  - We want to use X to predict Y
  - We only need the part in X that is informative about Y
  - We can squeeze X through a bottleneck formed by a limited set of codewords $\tilde{X}$, which preserves information about Y
  - There is a tradeoff between compressing the representation and preserving meaningful information



$$\min_{p(\tilde{x}|x)} \mathcal{L}[p(\tilde{x}|x)] = I(X;\tilde{X}) - \beta \cdot I(\tilde{X};Y\,)$$

Minimize Mutual Information between $\mathbf{X}$ and $\tilde{X}$ to achieve a minimal representation

Lagrange multiplier ß $> \mathbf{0}$ balances the constraints

Maximize Mutual Information between $\tilde{X}$ and Y to achieve a good prediction

N Tishby, F. Pereira, and W. Bialek (2000): The information bottleneck method. https://arxiv.org/pdf/physics/0004057.pdf Slide courtesy of H. Gupta

Summary

- Information Theory …
  - provides measures for information content of data, prior knowledge, models
  - is a general framework for system analysis and model building
  - is discipline-independent

www.victorinox.com

Further reading

- Papers
  - Shannon, C. E. (1948): A mathematical theory of communication, Bell system technical journal, 27, 623-656, citeulike-article-id:1584479. The big bang paper
  - Kullback, S., and Leibler, R. A. (1951): On Information and Sufficiency, Ann. Math. Statist., 22, 79-86, 10.1214/aoms/1177729694. Another classic
  - Weijs, S. (2011): Information Theory for Risk-based Water System Operation, PhD thesis, Faculty of Civil Engineering & Geosciences, section Water Resources Management, TU Delft, The Netherlands, Delft, 210 pp. Very accessible and applied
  - Pechlivanidis, I.G.; Jackson, B.; McMillan, H.; Gupta, H.V. (2016): Robust informational entropy-based descriptors of flow in catchment hydrology. Hydrol. Sci. J. , 61, 1–18. Robust binning
  - Knuth, K. H. (2019): Optimal data-based binning for histograms and histogram-based probability density models, Digital Signal Processing, 95, 102581. Optimal binning
  - Kumar, P., and Ruddell, B. L. (2010): Information Driven Ecohydrologic Self-Organization, Entropy, 12, 2085-2096, 10.3390/e12102085. Application to Ecohydrological context

Further reading

- Textbooks
  - Cover, T., and Thomas, J. A. (2006): Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience. Rigorous, comprehensive, the must-have
  - Singh, V. P. (213): Entropy Theory and its Application in Environmental and Water Engineering, John Wiley & Sons, Ltd. Various application examples

- Journals
  - Entropy (MDPI) https://www.mdpi.com/journal/entropy. All kinds of disciplines and applications

- Tutorials
  - Nicholas Timme and Christopher Lapish: A Tutorial for Information Theory in Neuroscience. https://www.eneuro.org/content/5/3/ENEURO.0052-18.2018. Great tutorial for newcomers

10001110100111111100000010111100111001010101000001111111111000000011010101010010