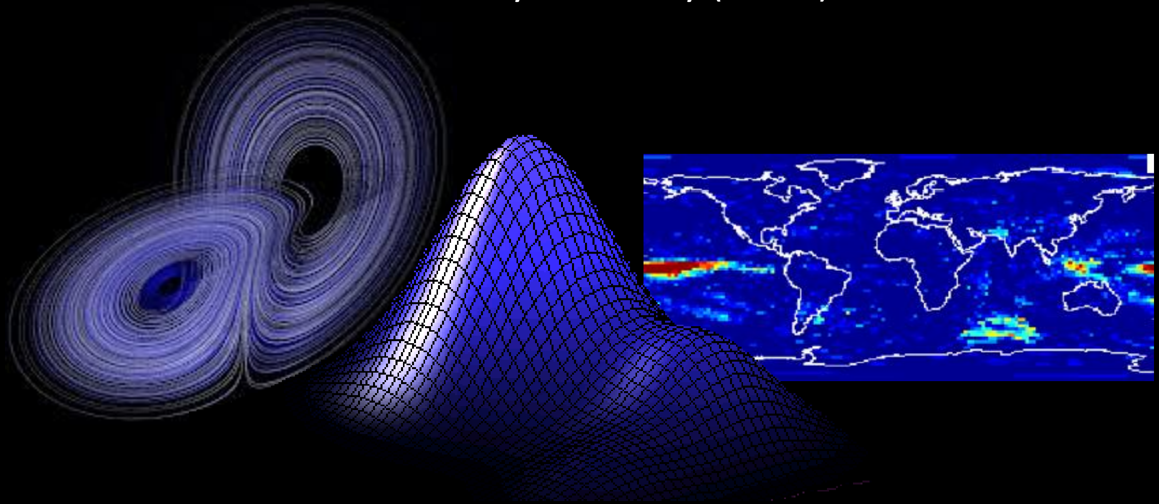
The Relationship Between Information and Physics

Kevin H. Knuth
Departments of Physics and Informatics
University at Albany (SUNY)



***Entropy* — Open Access Journal**

IF (2014): 1.502
IF (2015 projected): 1.715

Familiarity breeds the illusion of understanding

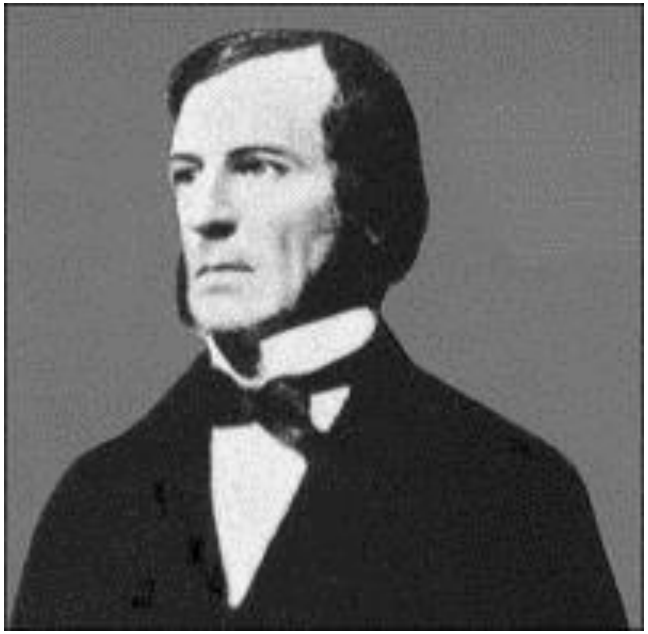
Anonymous

George Boole and Boolean Logic

George Boole was the inventor of Boolean Logic.

In 1854 he published:

“An Investigation of the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities”



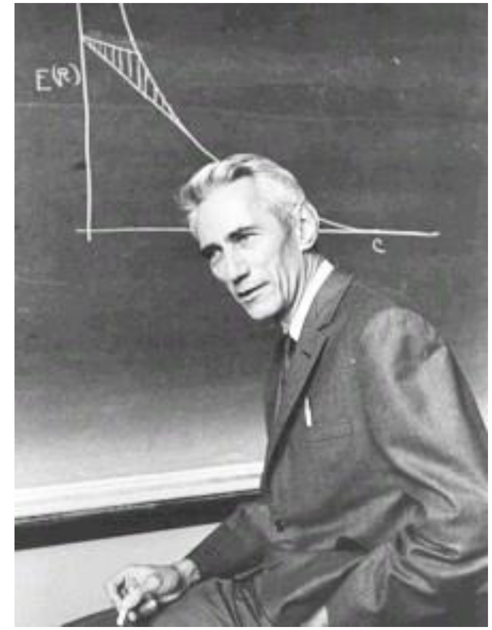
George Boole (1815 – 1874)

Since that time, HUNDREDS OF THOUSANDS of papers have been published on this topic.

Claude Shannon and Information Theory

In what is perhaps the most important Masters Thesis of the 20th Century,

Claude Shannon realized that Boolean logic could be used to optimize arrays of electromagnetic relays used in switching telephone systems.

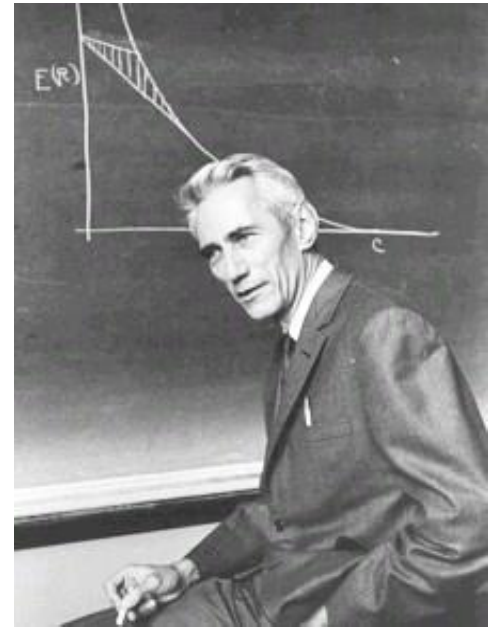


Claude Shannon (1916 – 2001)

With this insight that switches could emulate Boolean logic operations, we entered the Computer Age!

Claude Shannon and Information Theory

Years later at AT&T Bell Labs,
Claude Shannon derived a logically consistent way of quantifying the amount of information that could be transferred in a communication channel



Claude Shannon (1916 – 2001)

This resulted in the Shannon Entropy and Information Theory!

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log p(x)$$

Inferential Reasoning

Rev. Thomas Bayes
(1702 – 1761)

Richard T. Cox
(1898 – 1991)

Edwin T. Jaynes
(1922 – 1998)

The extension of Boolean logic to Bayesian Probability Theory extends the deductive logic of a traditional computer to inferential reasoning, which is capable of handling uncertainty.

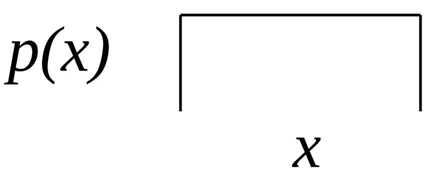
Expectation and Surprise!

Expectation and Surprise

When Henry was born, he had no information about the world.

All things were essentially equally probable.

He was equally surprised by everything.

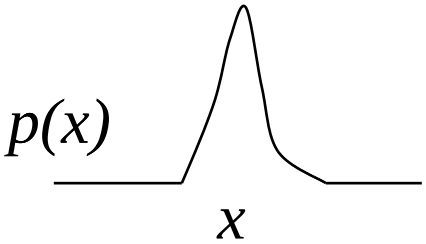


All states equally probable.

Expectation and Surprise

Henry now has some idea that some events are more probable than others.

He is now sometimes surprised!



Some states are common and others rarely occur!

Surprise

We use x to denote a particular state of the system out of a set of possible states X

The **surprise** is large for improbable states and small for probable states.

$$h(x) = \log \frac{1}{p(x)}$$

Surprise

The **surprise** is large for improbable states and small for probable states.

$$h(x) = \log \frac{1}{p(x)}$$

Averaging the surprise over all of the possible states of the system gives a measure of our uncertainty about the states of a system:

$$H(X) \;=\; \sum_{x \in X} p(x) \log \frac{1}{p(x)} \;=\; -\sum_{x \in X} p(x) \log p(x)$$

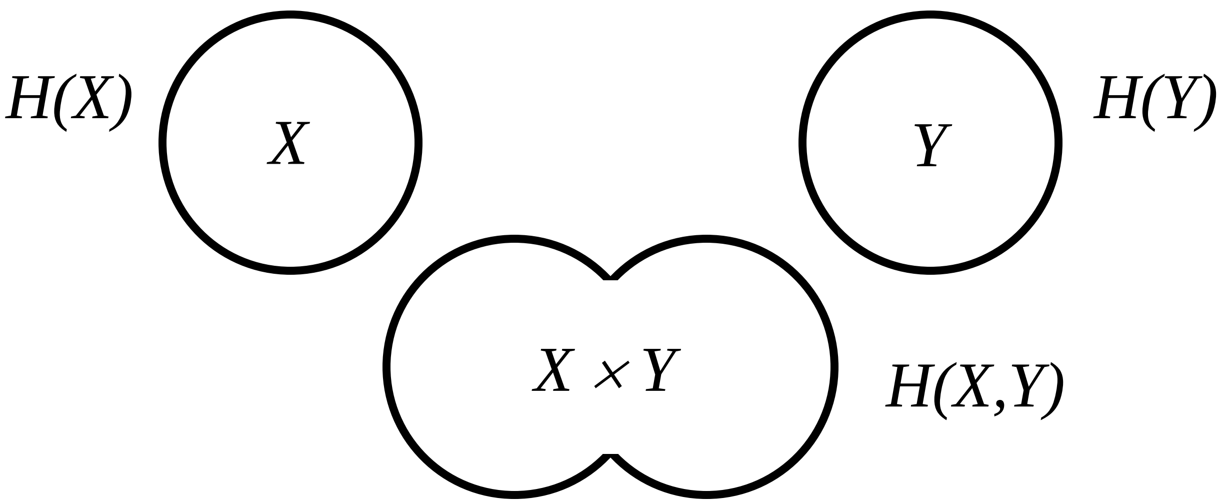
which is called the Shannon **entropy**.

Joint Entropy

If the system states can be described with multiple parameters, the entropy is computed by averaging over all possible states

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

This is called the **Joint Entropy**, since it describes the entropy of the states of X and Y, which jointly describe the system. You can think of X and Y as representing subsystems of the original system $X\times Y$.

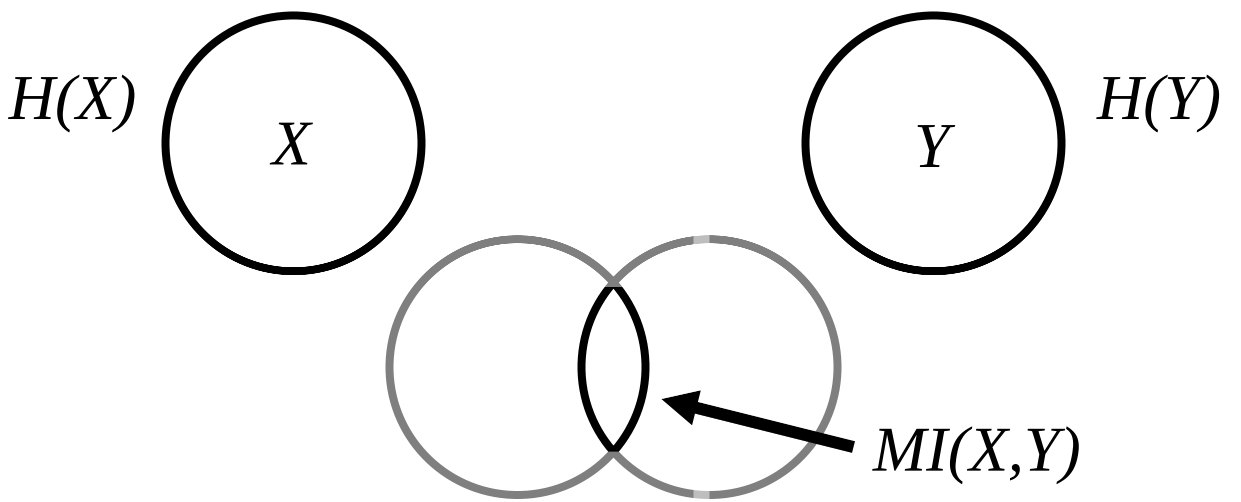


Mutual Information

One can consider a joint system which is composed from joining two systems. In this case, an important quantity is the difference of entropies,

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$

This is called the **Mutual Information** (MI) since it describes the amount of information that is shared between the two subsystems.



The Laws of Nature

From Where do the Laws of Nature Originate?

From Where do the Laws of Nature Originate?

Laws are fundamental and are dictated by God or Mother Nature or Historical Accident

It is widely believed that the Laws of Nature reflect underlying order in the universe

From Where do the Laws of Nature Originate?

Laws are fundamental and are dictated by God or Mother Nature or Historical Accident

Laws are based on fundamental symmetries

Laws reflect the optimal means by which one can process information about the universe

From Where do the Laws of Nature Originate?

Dictated ⇨ *implies they must be discovered*

Symmetries ⇨ *laws are relations enforcing symmetries*

Optimal Information Processing ⇨ *probability and entropy*

From Where do the Laws of Nature Originate?

Dictated ⇨ *implies they must be discovered*

Symmetries ⇨ *laws are relations enforcing symmetries*

Optimal Information Processing ⇨ *probability and entropy*

These latter two paradigms imply that Laws might be derived!

Information Physics

Symmetries

Consistent Quantification

Cox, Jaynes, Knuth and Skilling
Quantification of Statements

Knuth
Quantification of Questions

Goyal, Knuth, Skilling
Quantum Mechanics
(Quantified Measurement Sequences)

Knuth, Bahreyni and Walsh
Special Relativity
Spacetime Physics
Relativistic Quantum Mechanics

Optimal Information Processing

Physics as Inference

Jaynes
Statistical Mechanics as Inference
Maximum Entropy

Caticha, Johnson, Cafaro, Nawaz, Abedi, Ipek, Bartolomeo, etc.
Entropic Dynamics

Dewar, Lorenz, Martyushev, Wang, etc
Maximum Entropy Production

Edwin T. Jaynes
(1922 – 1998)

Shortly after Shannon's work on Information Theory, Ed Jaynes realized that the Shannon Entropy was the same quantity as the entropy in statistical mechanics. This led to the development of:

The Principle of Maximum Entropy

where one assigns probabilities that maximize the entropy subject to any known constraints. In this sense statistical mechanics is a theory of inference.

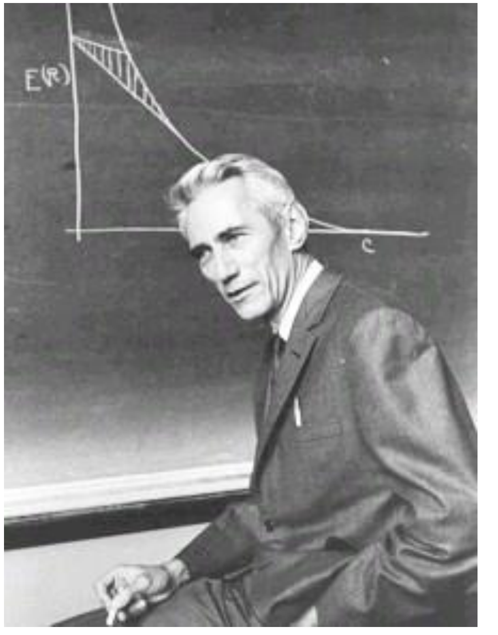
Edwin T. Jaynes
(1922 – 1998)

The idea behind
The Principle of Maximum Entropy
is to assign probabilities that are consistent with what is known, but are maximally ignorant otherwise (thereby not accidentally assuming something inappropriate)

Maximum entropy production is a similar concept applied to dynamical systems.

To paraphrase an analogy made by Brendon Brewer:

If it is true that many ways lead to the summit, then if you are on a path, you will very likely find your way to the summit!



It was a matter of great debate from the 1950s through the 2000s as to whether Information Theory was applicable to a wider array of problems than the communication channels for which it was developed.

Shannon weighed in on this debate stating that he did not believe that Information Theory was applicable outside of communication channels.

Fundamental Questions

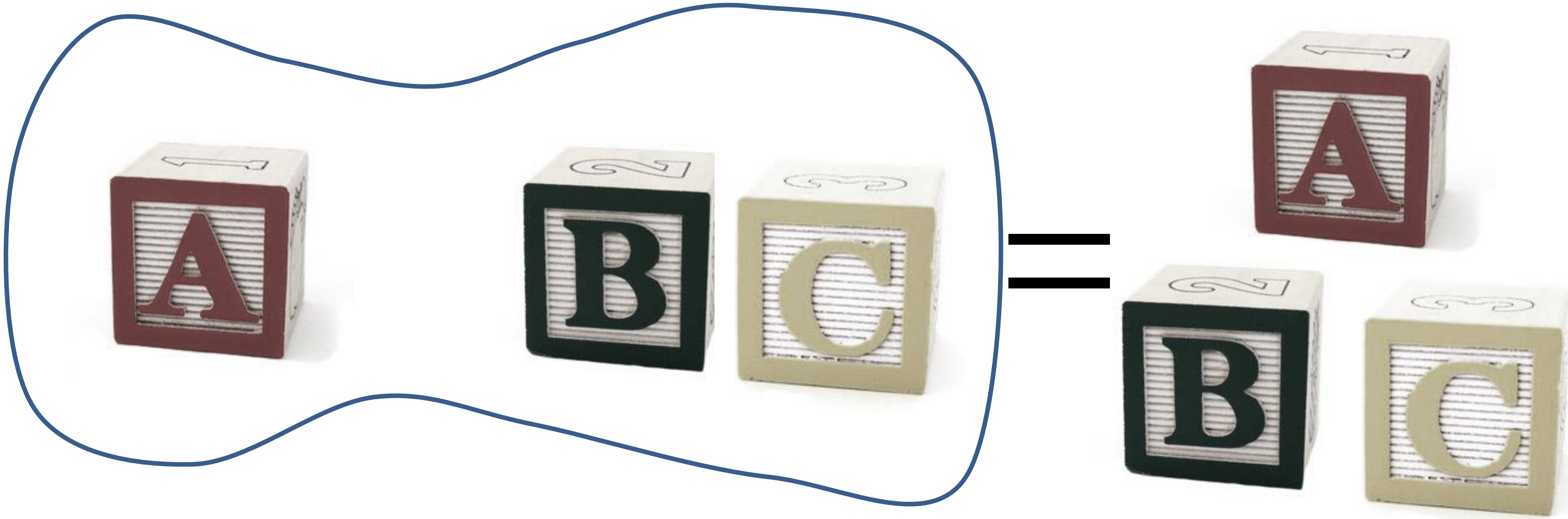
In graduate school I asked:

Why when I combine one crayon with two crayons

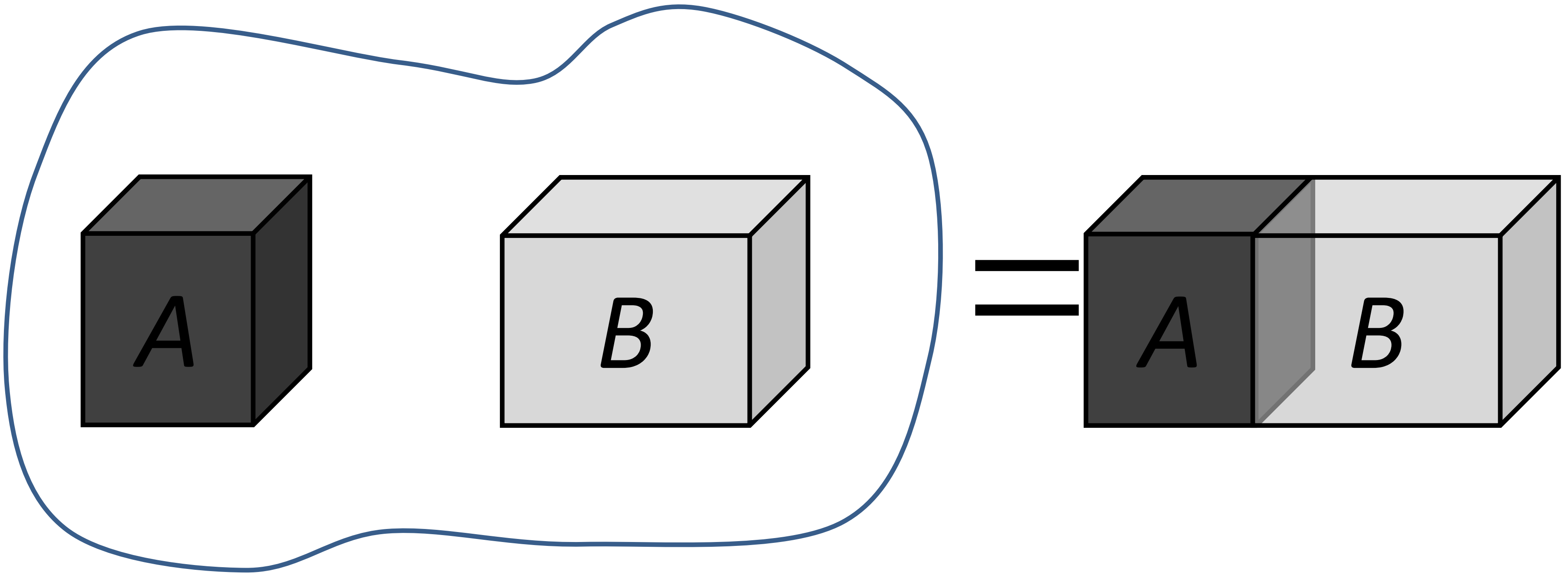
do I always get three crayons

$$1 + 2 = 3$$

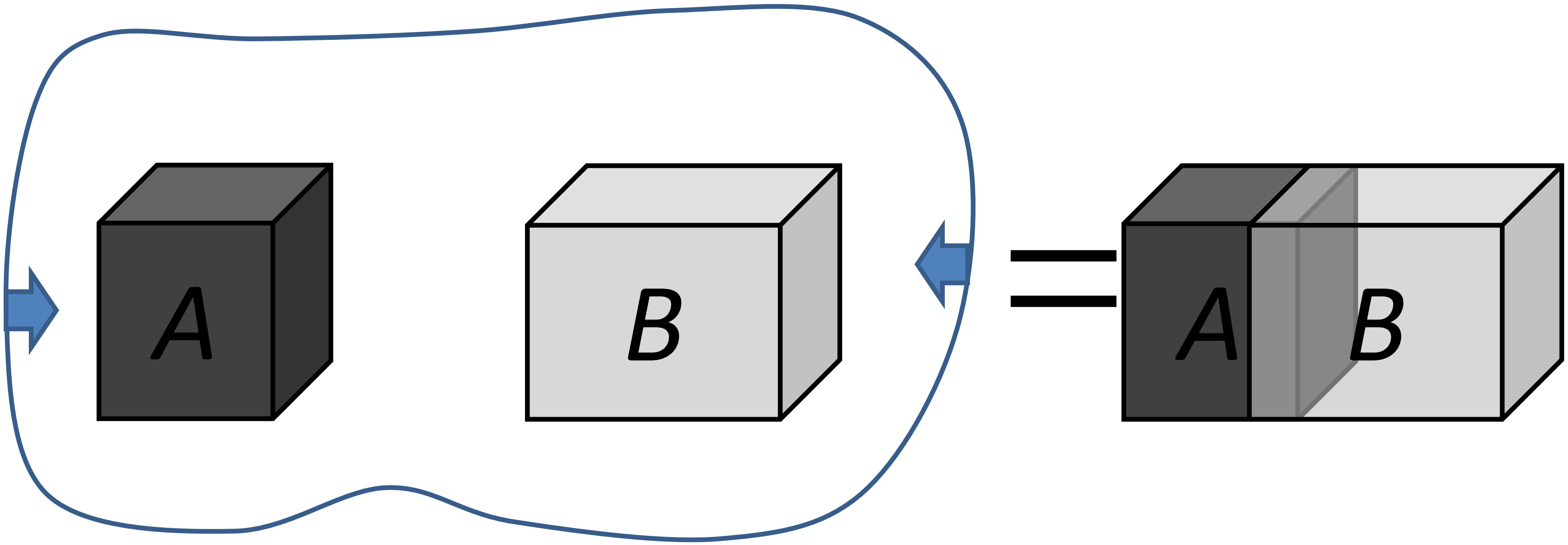
$$1 + 2 = 3$$



$$1 + 2 = 3$$



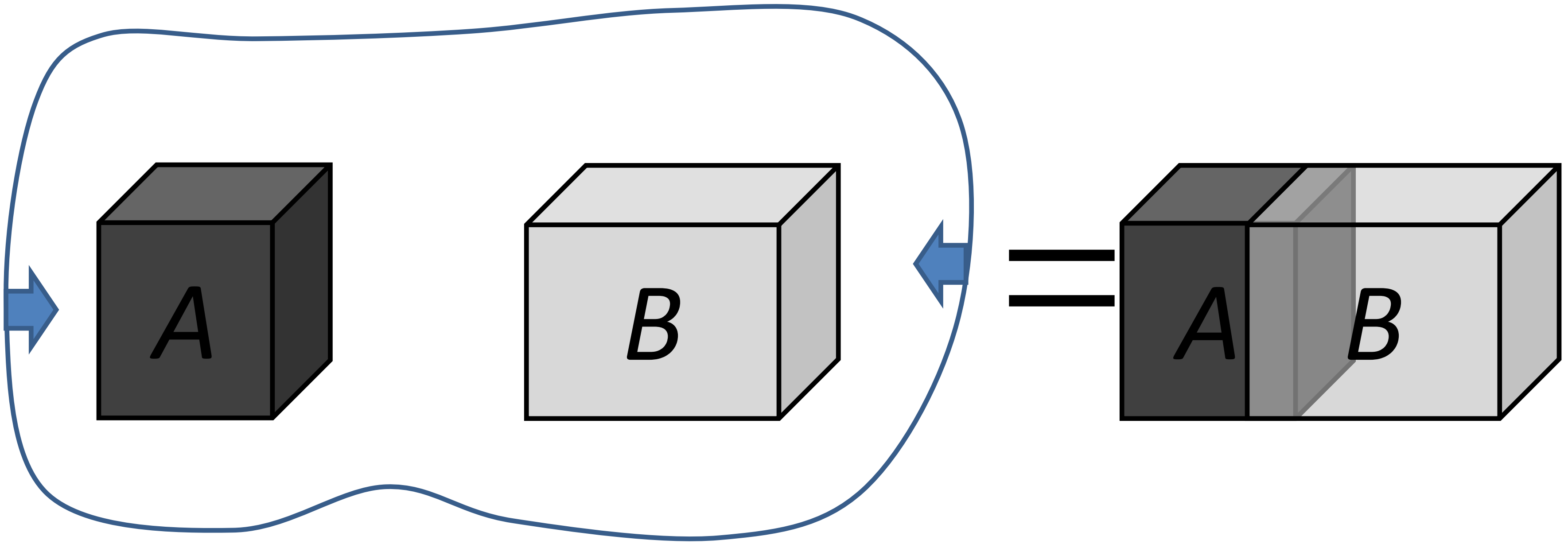
$$v(A \cup B) = v(A) + v(B)$$



$$v(A \cup B) = v(A) + v(B) - v(A \cap B)$$

volume

Knuth, MaxEnt 2003



$$s(A \cup B) = s(A) + s(B) - s(A \cap B)$$

surface area

Knuth, MaxEnt 2003

$$\Pr(A \vee B \mid I) = \Pr(A \mid I) + \Pr(B \mid I) - \Pr(A \wedge B \mid I)$$

sum rule of probability

Knuth, MaxEnt 2003

$$I(A;B) = H(A) + H(B) - H(A,B)$$

mutual information

Knuth, MaxEnt 2003

$$max(a,b) = a + b - min(a,b)$$

polya's min-max rule

Knuth, MaxEnt 2003

$$\log(LCM(a,b)) = \log(a) + \log(b) - \log(GCD(a,b))$$

number theory identity

Knuth, MaxEnt 2009

Clearly, my original question at to why

results in

$$1 + 2 = 3$$

Is related to many problems, but specifically this one:

why the probability of the disjunction of two statements A and B given I results in

$$\Pr(A \vee B \mid I) = \Pr(A \mid I) + \Pr(B \mid I) - \Pr(A \wedge B \mid I)$$

the essential content of both statistical mechanics and communication theory, of course, does not lie in the equations; it lies in the ideas that lead to those equations

E. T. Jaynes

the essential content of both statistical mechanics and communication theory, of course, does not lie in the equations; it lies in the ideas that lead to those equations

E. T. Jaynes

the essential content of both statistical mechanics and communication theory, of course, does not lie in the equations; it lies in the ideas that lead to those equations

E. T. Jaynes

A MODERN PERSPECTIVE

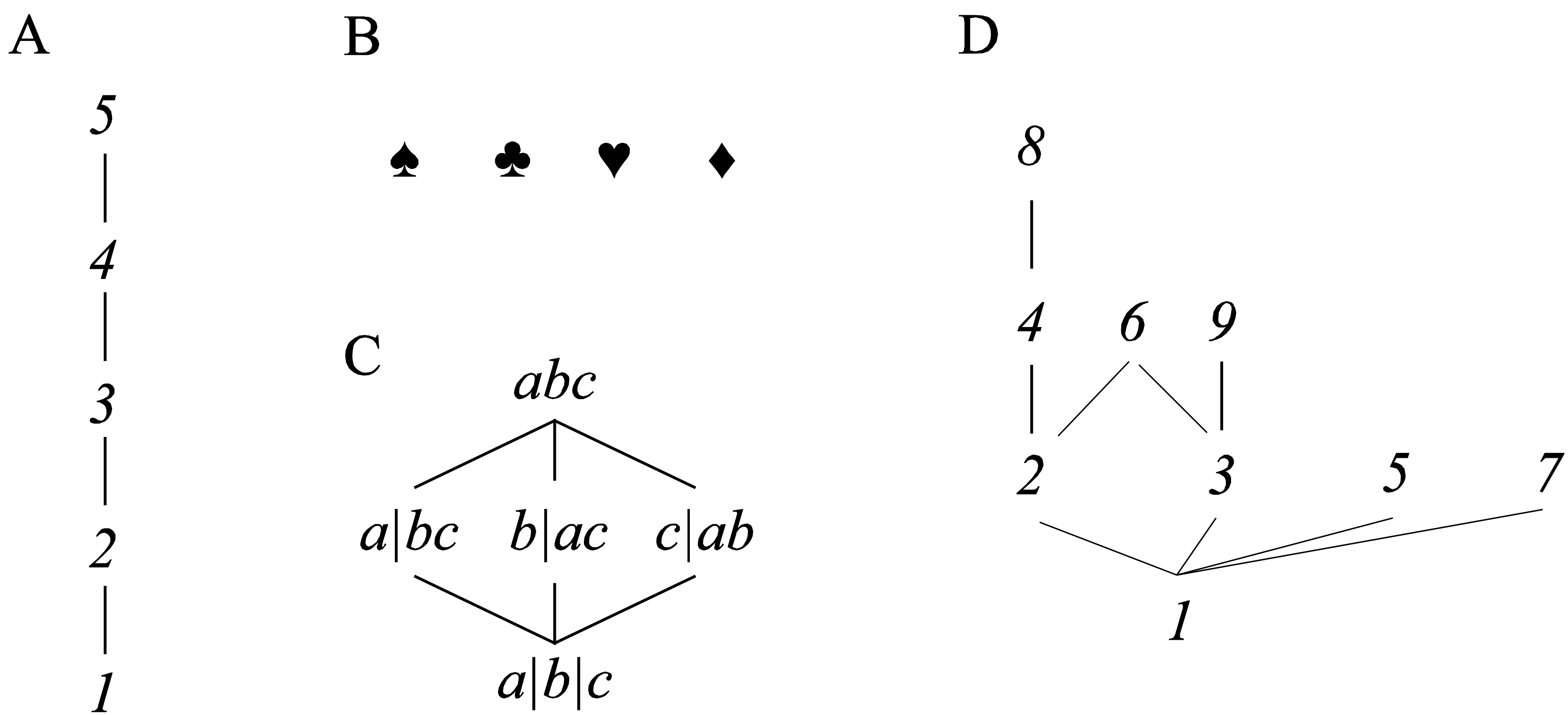
Measure what is measurable,

and make measurable that which is not so.

Galileo Galilei

Lattices

Lattices are partially ordered sets where each pair of elements has a least upper bound and a greatest lower bound



Lattices are Algebras

Structural Viewpoint		*Operational Viewpoint*
$a \le b$	$\Leftrightarrow$	$a \vee b = b$ $a \wedge b = a$

Lattices

Structural Viewpoint | *Operational Viewpoint*

$$a \le b \quad \Leftrightarrow \quad \begin{matrix} a \vee b = b \\ a \wedge b = a \end{matrix}$$

Sets, Is a subset of

$$a \subseteq b \quad \Leftrightarrow \quad \begin{matrix} a \cup b = b \\ a \cap b = a \end{matrix}$$

Positive Integers, Divides

$$a \mid b \quad \Leftrightarrow \quad \begin{matrix} \mathrm{lcm}(a,b) = b \\ \gcd(a,b) = a \end{matrix}$$

Assertions, Implies

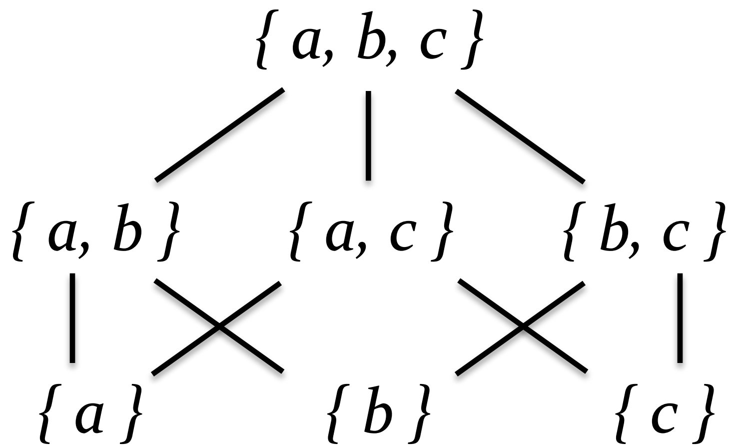
$$a \to b \quad \Leftrightarrow \quad \begin{matrix} a \vee b = b \\ a \wedge b = a \end{matrix}$$

Integers, Is less than or equal to

$$a \le b \quad \Leftrightarrow \quad \begin{matrix} \max(a,b) = b \\ \min(a,b) = a \end{matrix}$$

Quantification

quantify the partial order $\equiv$ *assign real numbers to the elements*

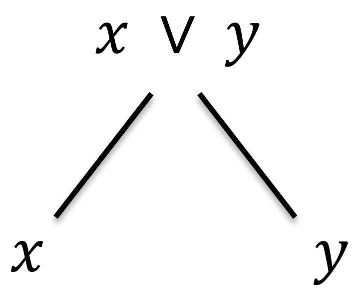


$$f : x \in L \rightarrow \mathbb{R}$$

Require that quantification be consistent with the structure.
Otherwise, information about the partial order is lost.

Local Consistency

Any general rule must hold for special cases
Look at special cases to constrain general rule



$f: x \in L \rightarrow \mathbb{R}$

Enforce local consistency

$$f(x \vee y) = f(x) \oplus f(y)$$

where $\oplus$ is an unknown operator to be determined.

Associativity of Join

Write the same element two different ways

$$x \vee (y \vee z) = (x \vee y) \vee z$$

which implies

$$f(x) \oplus \big(f(y) \oplus f(z)\big) = \big(f(x) \oplus f(y)\big) \oplus f(z)$$

Note that the unknown operator $\oplus$ is nested in two distinct ways, which reflects associativity

Associativity Equation

This is a functional equation known as the
Associativity Equation

$$f(x) \oplus \left(f(y) \oplus f(z)\right) = \left(f(x) \oplus f(y)\right) \oplus f(z)$$

where the aim is to find all the possible operators $\oplus$ that satisfy the equation above.

We require that the join operations are closed,
That the valuations respect ranking, i.e. $x \geq y \Rightarrow f(x) \geq f(y)$
And that $\oplus$ is commutative and associative.

Associativity Equation

The general solution to the Associativity Equation

$$f(x) \oplus \big(f(y) \oplus f(z)\big) = \big(f(x) \oplus f(y)\big) \oplus f(z)$$

is (Aczel 1966; Craigen and Pales 1989; Knuth and Skilling 2012):

$$F\big(f(x) \oplus f(y)\big) = F\big(f(x)\big) + F\big(f(y)\big)$$

where F is an arbitrary invertible function.

Regraduation

$$F\big(f(x)\ \oplus f(y)\big) = F\big(f(x)\big) + F\big(f(y)\big)$$

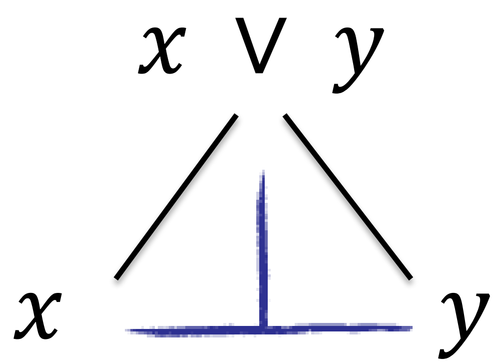
Since the function F is arbitrary and invertible, we can define a new quantification $v(x) = F\big(f(x)\big)$ so that the combination is always additive.

Thus we can always write

$$v(x \vee y) = v(x) + v(y)$$

*In essence, we have **derived measure theory** from algebraic symmetries.*

Additivity



$$v(x \vee y) = v(x) + v(y)$$

Knuth, MaxEnt 2009

Epiphany!

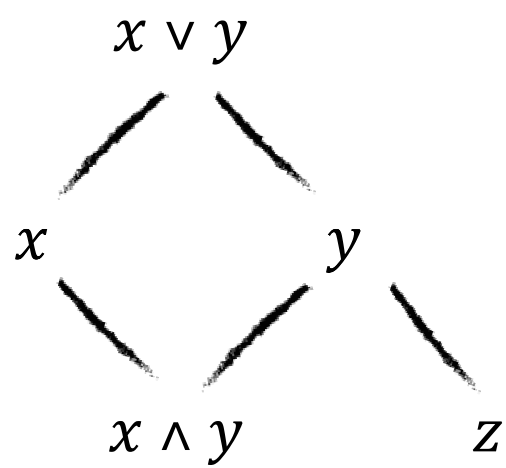
Why We Sum

always results in

$$1 + 2 = 3$$

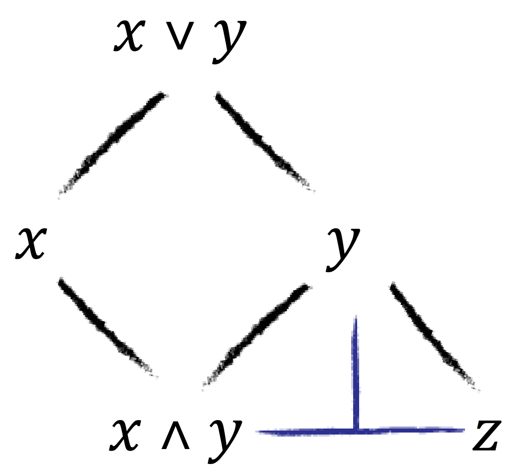
because it is guaranteed to always work since combining crayons in this way is closed, commutative, associative, and I can order sets of crayons.

General Case



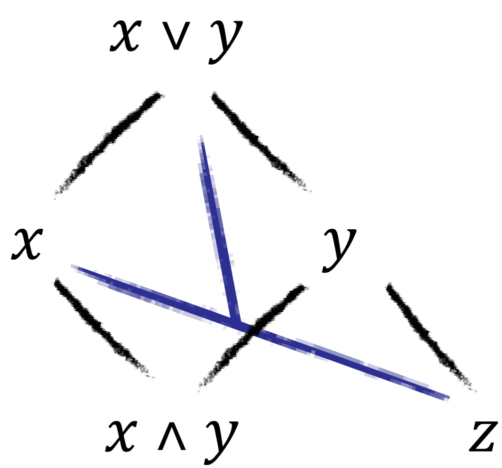
A More General Case

General Case



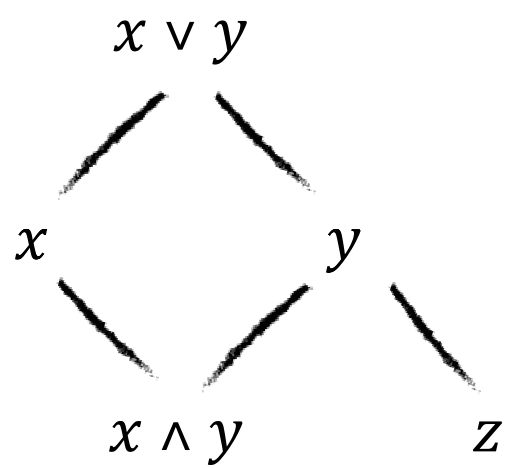
$$v(y) = v(x \wedge y) + v(z)$$

General Case



$$v(y) = v(x \wedge y) + v(z) \qquad v(x \vee y) = v(x) + v(z)$$

General Case

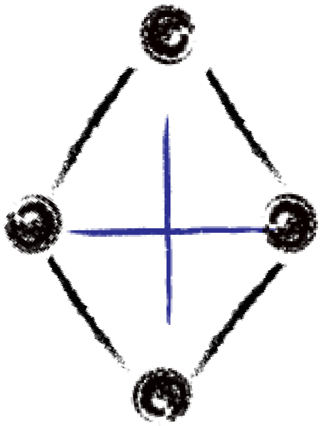


$$v(y) = v(x \wedge y) + v(z) \qquad v(x \vee y) = v(x) + v(z)$$

$$v(x \vee y) = v(x) + v(y) - v(x \wedge y)$$

Sum Rule

$$v(x \vee y) = v(x) + v(y) - v(x \wedge y)$$



$$v(x \vee y) + v(x \wedge y) = v(x) + v(y)$$

symmetric form (self-dual)

A Curious Observation

Fundamental symmetries are why the Sum Rule is ubiquitous

Ubiquity (inclusion-exclusion)

$\Pr(A \vee B \mid C) = \Pr(A \mid C) + \Pr(B \mid C) - \Pr(A \wedge B \mid C)$	*Probability*
$I(A;B) = H(A) + H(B) - H(A,B)$	*Mutual Information*
$Area(A \cup B) = Area(A) + Area(B) - Area(A \cap B)$	*Areas of Sets*
$\max(A,B) = A + B - \min(A,B)$	*Polya's Min-Max Rule*
$\log LCM(A,B) = \log A + \log B - \log GCD(A,B)$	*Integral Divisors*
$I_3(A,B,C) = \lvert A \sqcup B \sqcup C \rvert - \lvert A \sqcup B \rvert - \lvert A \sqcup C \rvert - \lvert B \sqcup C \rvert + \lvert A \rvert + \lvert B \rvert + \lvert C \rvert$	*Amplitudes from three-slits (Sorkin arXiv:\\gr-qc/9401003)*

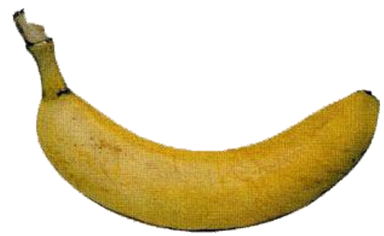
The relations above are constraint equations ensuring consistent quantification in the face of certain symmetries (commutativity, Associativity, Closure, and Ranking)

Knuth, 2003. Deriving Laws, arXiv:physics/0403031 [physics.data-an]
Knuth, 2009. Measuring on Lattices, arXiv:0909.3684 [math.GM]
Knuth, 2015. The Deeper Roles of Mathematics in Physical Laws, arXiv:1504.06686 [math.HO]

INFERENCE

What can be said about a system?

states

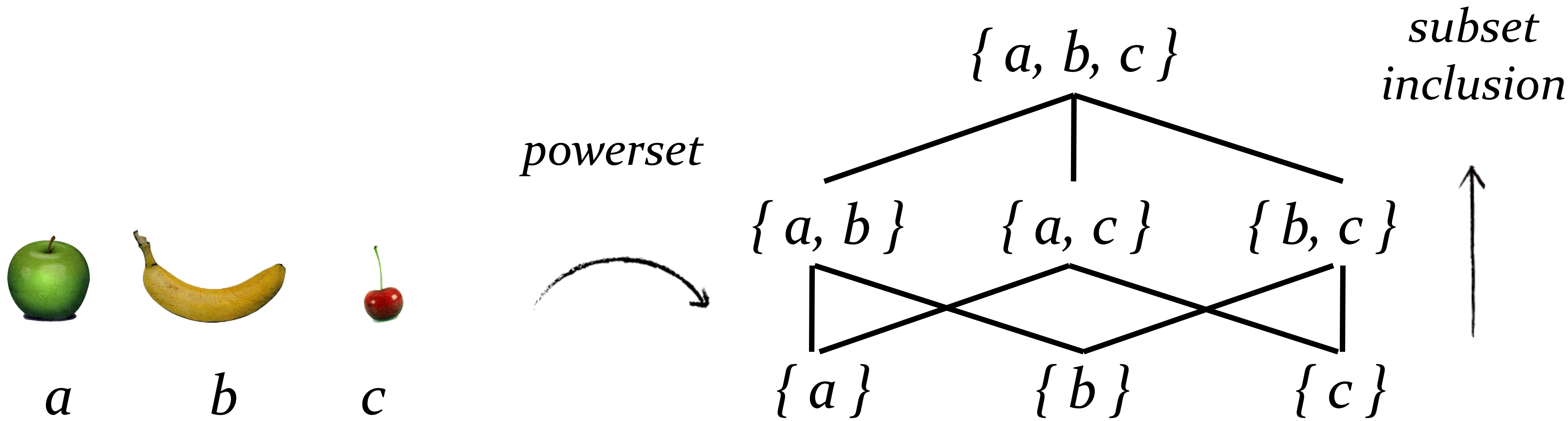


apple *banana* *cherry*

states of the contents of my grocery basket

What can be said about a system?

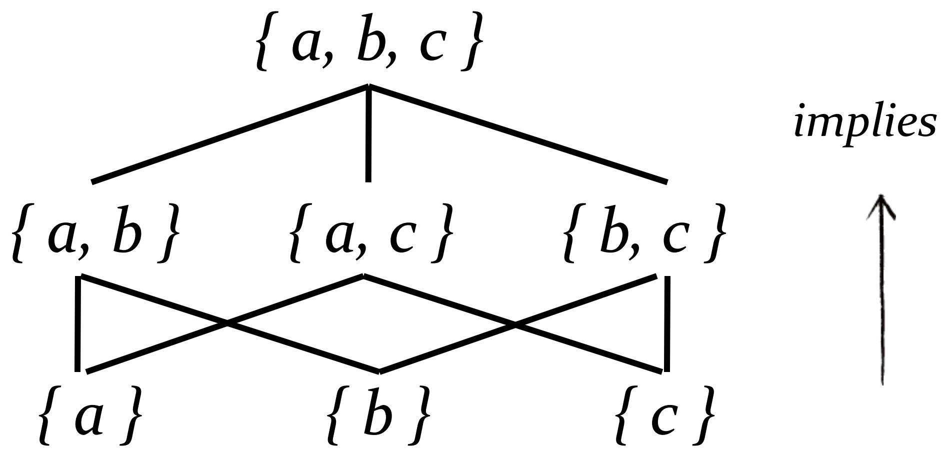
crudely describe knowledge by listing a set of potential states



states of the contents of my grocery basket

statements about the contents of my grocery basket

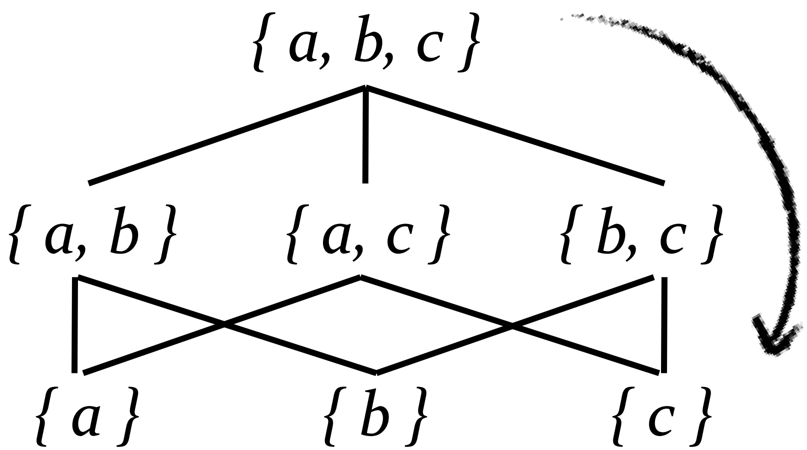
What can be said about a system?



statements
about the contents of
my grocery basket

ordering encodes implication
DEDUCTION

What can be said about a system?

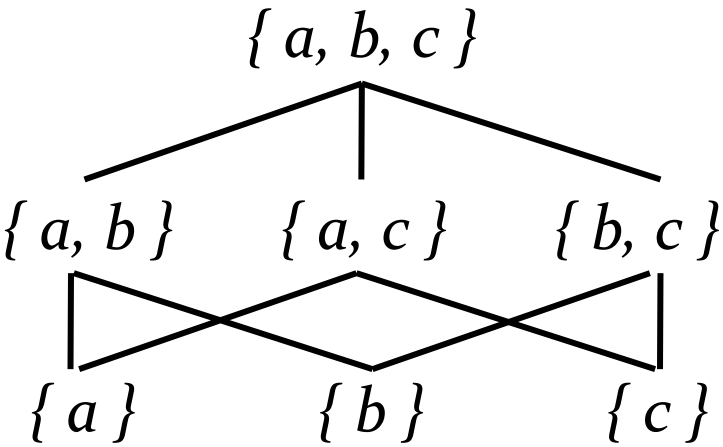


statements about the contents of my grocery basket

Quantify to what degree the statement that the system is in one of three states {a, b, c} implies knowing that it is in some other set of states

inference works backwards

Inclusion and the Zeta Function



The Zeta function encodes inclusion (Boolean implication) on the lattice.

$$\zeta(x,y) = \begin{cases} 1 & \text{if } x \leq y \\ 0 & \text{if } x \not\leq y \end{cases}$$

One can conceive of probability as a generalization of the zeta function (Boolean implicatio)

Context and Bi-Valuations

BI-VALUATION $p : x, i \in L \rightarrow R$

Bi-Valuation				*Valuation*
$p(x \mid i)$	→	$v_i(x)$	→	$v(x)$
Context i is explicit		*Measure of x with respect to Context i*		*Context i is implicit*

Bi-valuations generalize lattice inclusion to degrees of inclusion

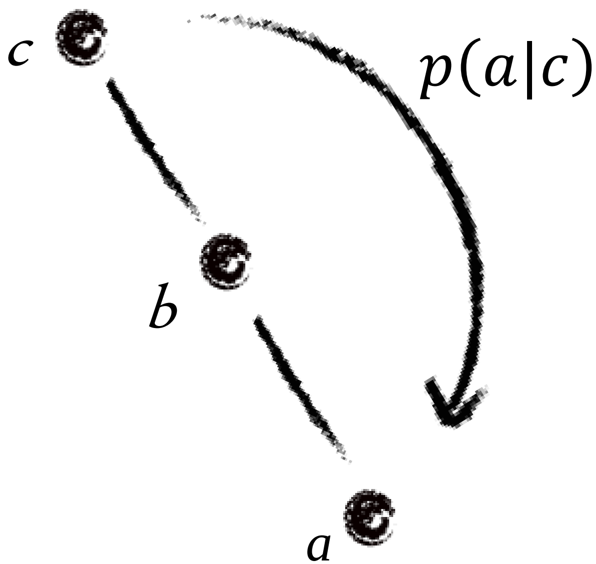
The logical disjunction (OR), ∨, *is associative, commutative, and closed.*

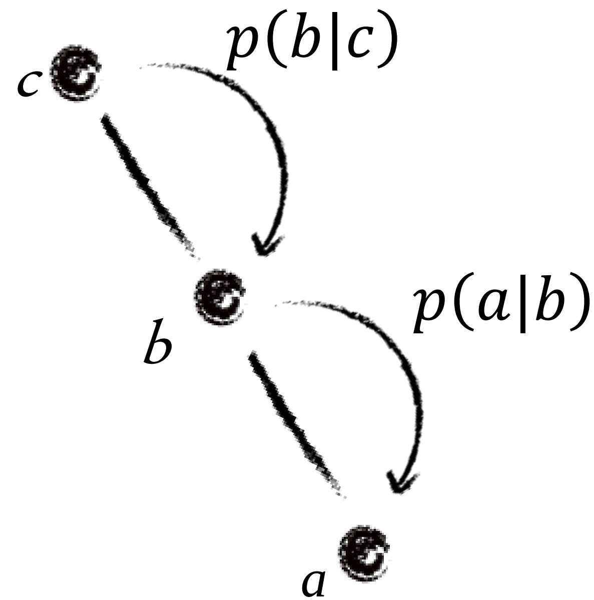
As a result, the valuations obey the ***Sum Rule*** *under constant context, i.*

$$p(x \mid i) + p(y \mid i) = p(x \vee y \mid i) + p(x \wedge y \mid i)$$

Changing Context

Context

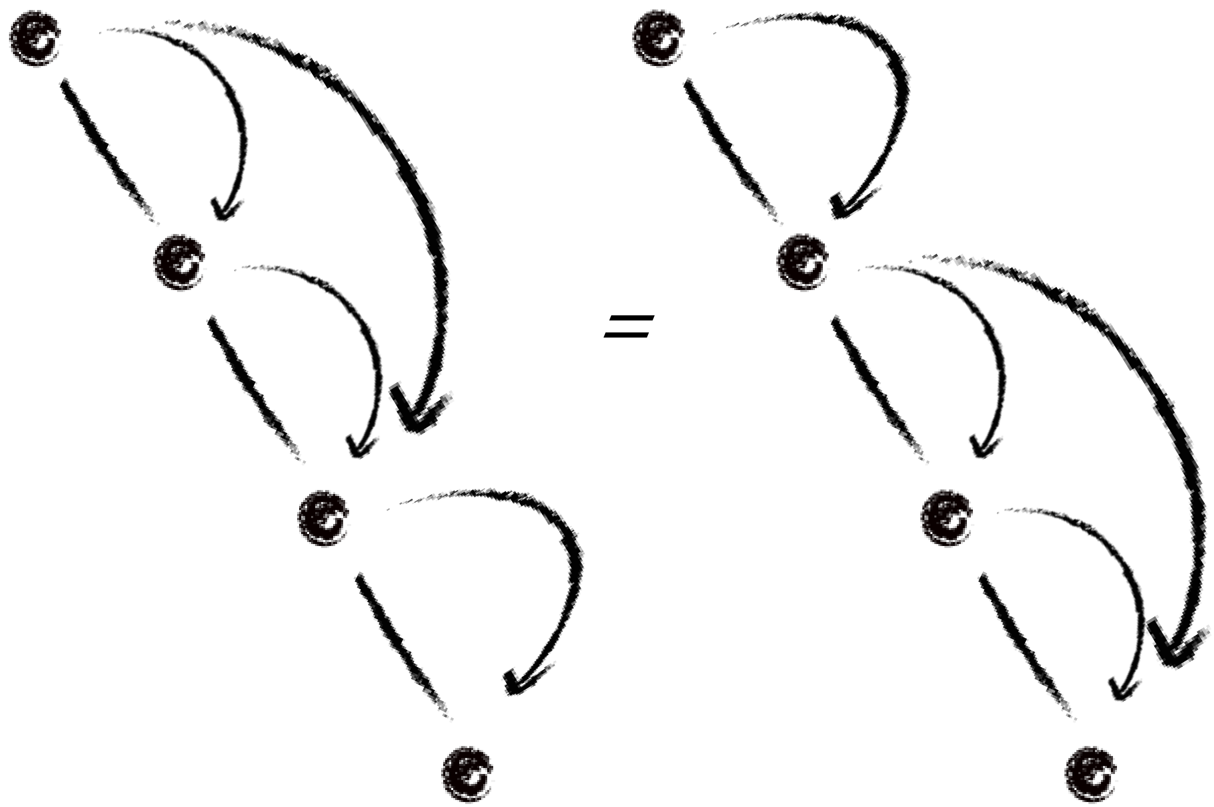




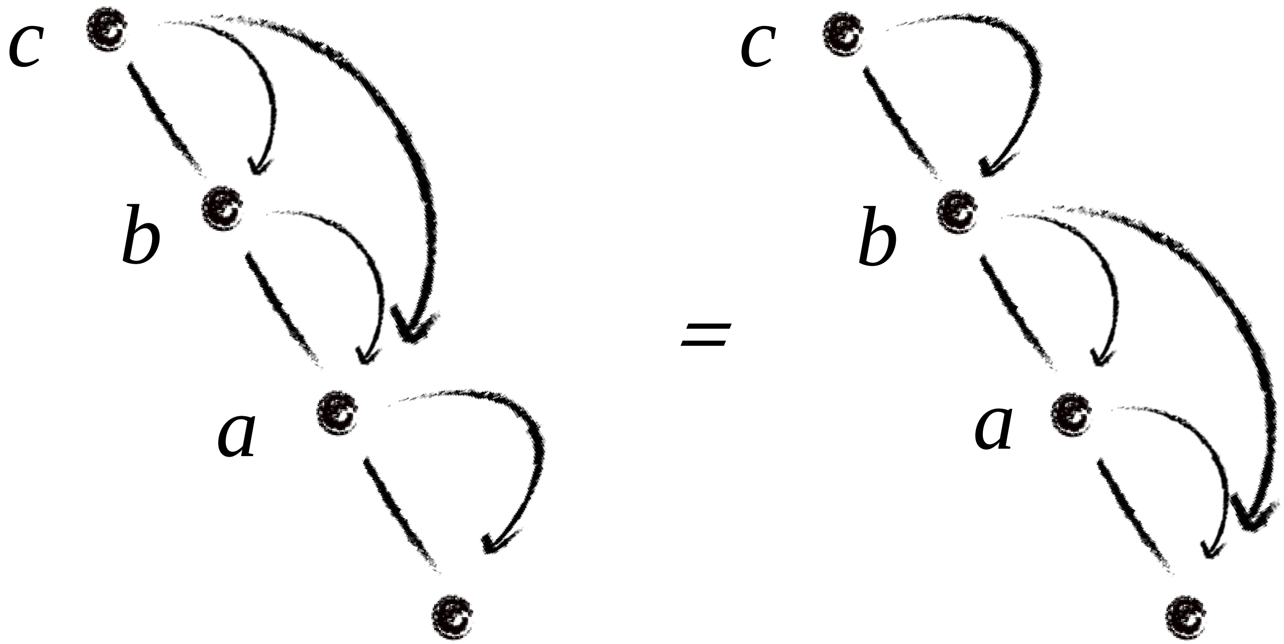
$$p(a|c) = p(a|b) \otimes p(b|c)$$

where the operator $\otimes$ is to be determined

Associativity of Context

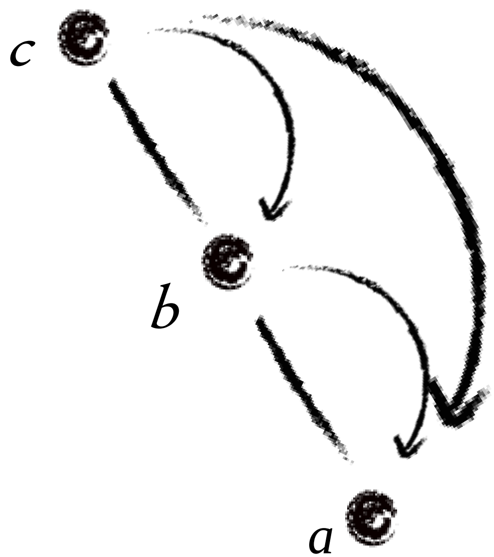


Associativity of Context



Since ⨂ is associative, commutative, and obeys closure, it must be an invertible transform of addition. However, the only degree of freedom left is that of scale so it must be a product.

Chain Rule



Chain Rule

$$p(a|c) = p(a|b)p(b|c)$$

How is the above an invertible transform of addivity?

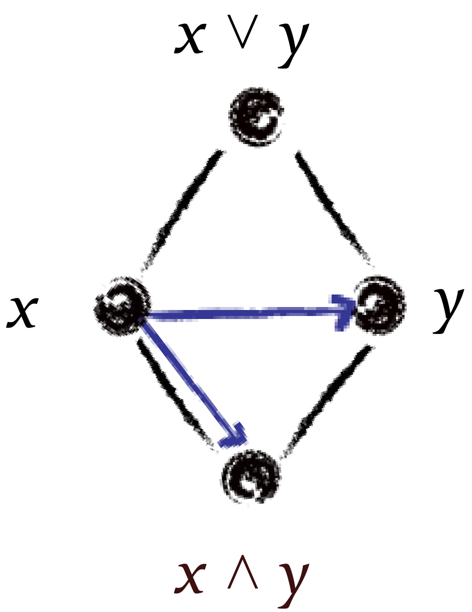
$$\log\big(p(a|c)\big) = \log\big(p(a|b)\big) + \log(p(b|c))$$

An Identity

Lemma

$$p(x \mid x) + p(y \mid x) = p(x \vee y \mid x) + p(x \wedge y \mid x)$$

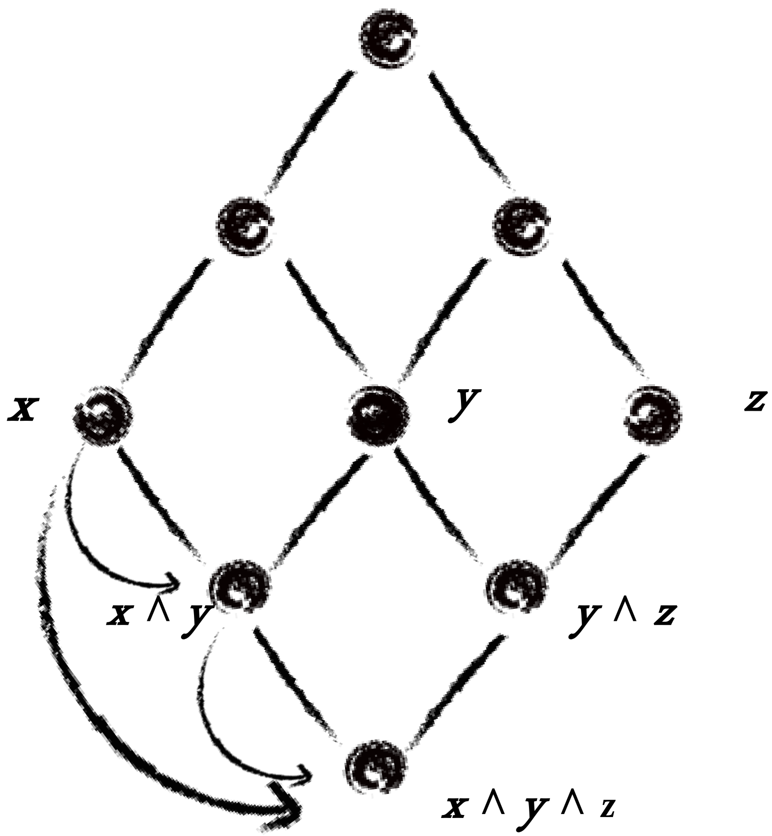
Since $x \leq x$ *and* $x \leq x \vee y$, $p(x \mid x) = 1$ *and* $p(x \vee y \mid x) = 1$



$$p(y \mid x) = p(x \wedge y \mid x)$$

Extending the Chain Rule

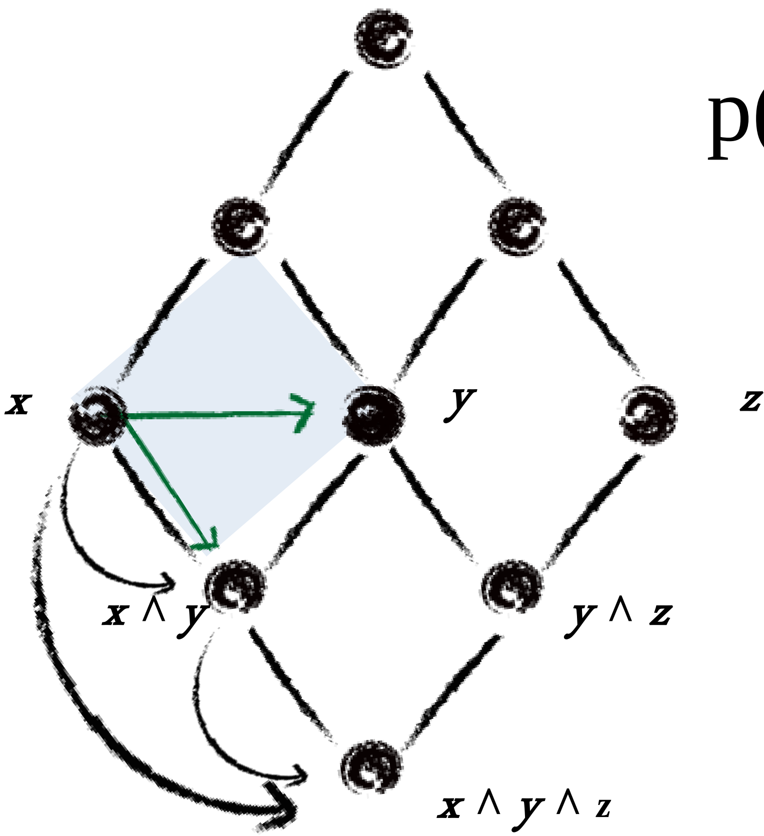
$$p(x \wedge y \wedge z \mid x) = p(x \wedge y \mid x)\, p(x \wedge y \wedge z \mid x \wedge y)$$



Extending the Chain Rule

$$p(x \wedge y \wedge z \mid x) = p(x \wedge y \mid x)\, p(x \wedge y \wedge z \mid x \wedge y)$$

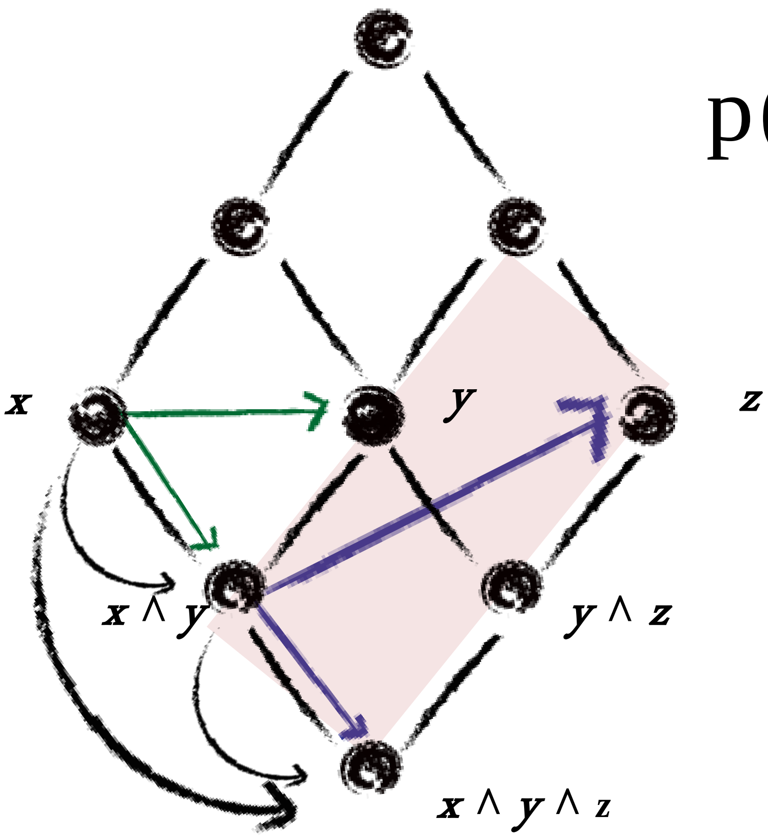
$$p(y \wedge z \mid x) = p(y \mid x)\, p(z \mid x \wedge y)$$



Extending the Chain Rule

$$p(x \wedge y \wedge z \mid x) = p(x \wedge y \mid x)\, p(x \wedge y \wedge z \mid x \wedge y)$$

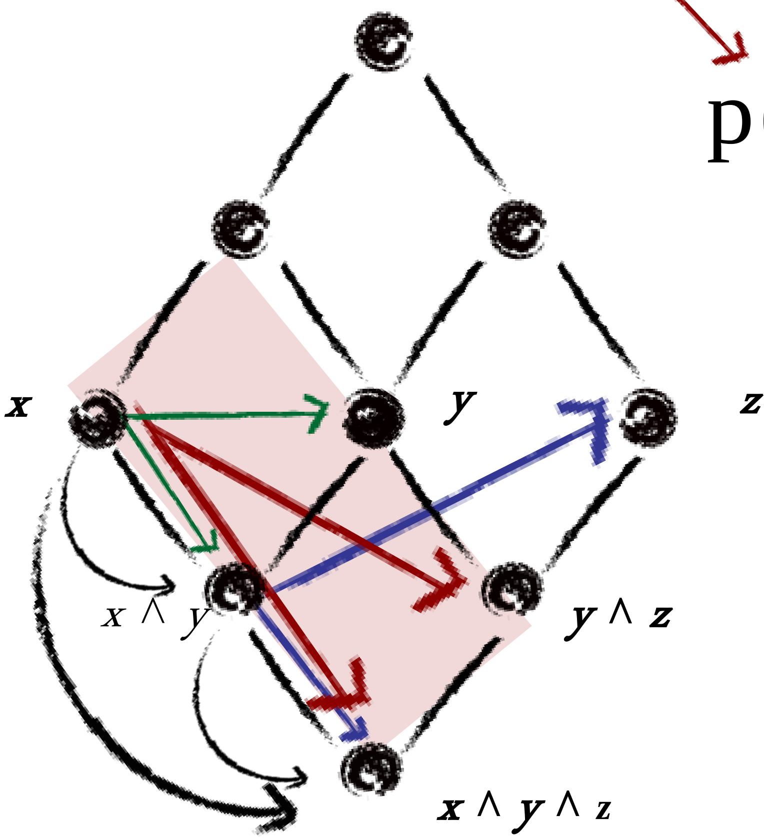
$$p(y \wedge z \mid x) = p(y \mid x)\, p(z \mid x \wedge y)$$



Extending the Chain Rule

$$p(x \wedge y \wedge z \mid x) = p(x \wedge y \mid x)\, p(x \wedge y \wedge z \mid x \wedge y)$$

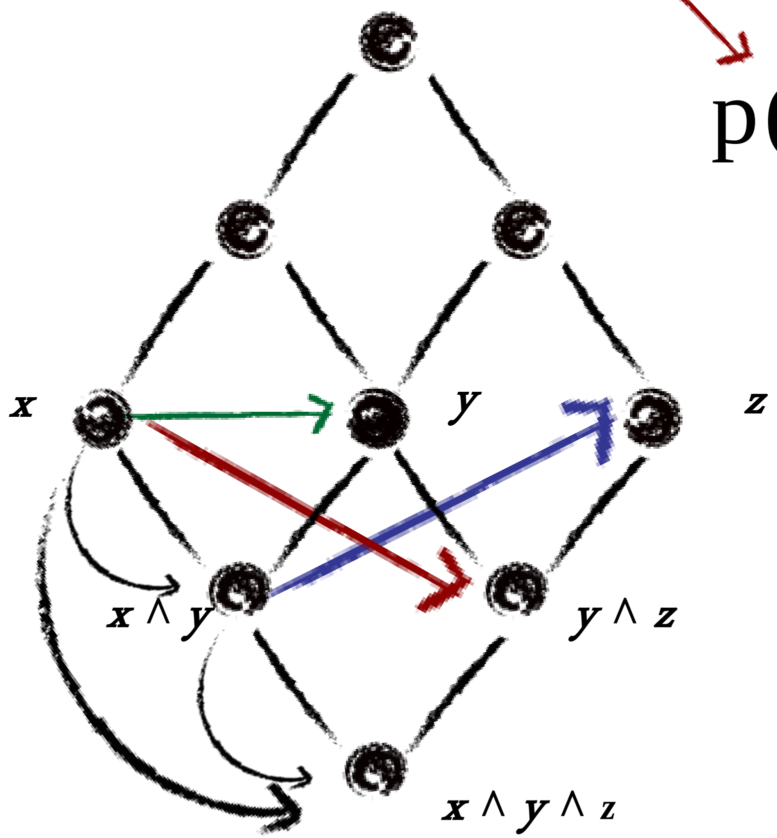
$$p(y \wedge z \mid x) = p(y \mid x)\, p(z \mid x \wedge y)$$



The Product Rule

$$p(x \wedge y \wedge z \mid x) = p(x \wedge y \mid x)\, p(x \wedge y \wedge z \mid x \wedge y)$$

$$p(y \wedge z \mid x) = p(y \mid x)\, p(z \mid x \wedge y)$$



Which is the familiar ***Product Rule****!*

Bayes Theorem and Change of Context

Commutativity of the product
leads to ***Bayes Theorem…***

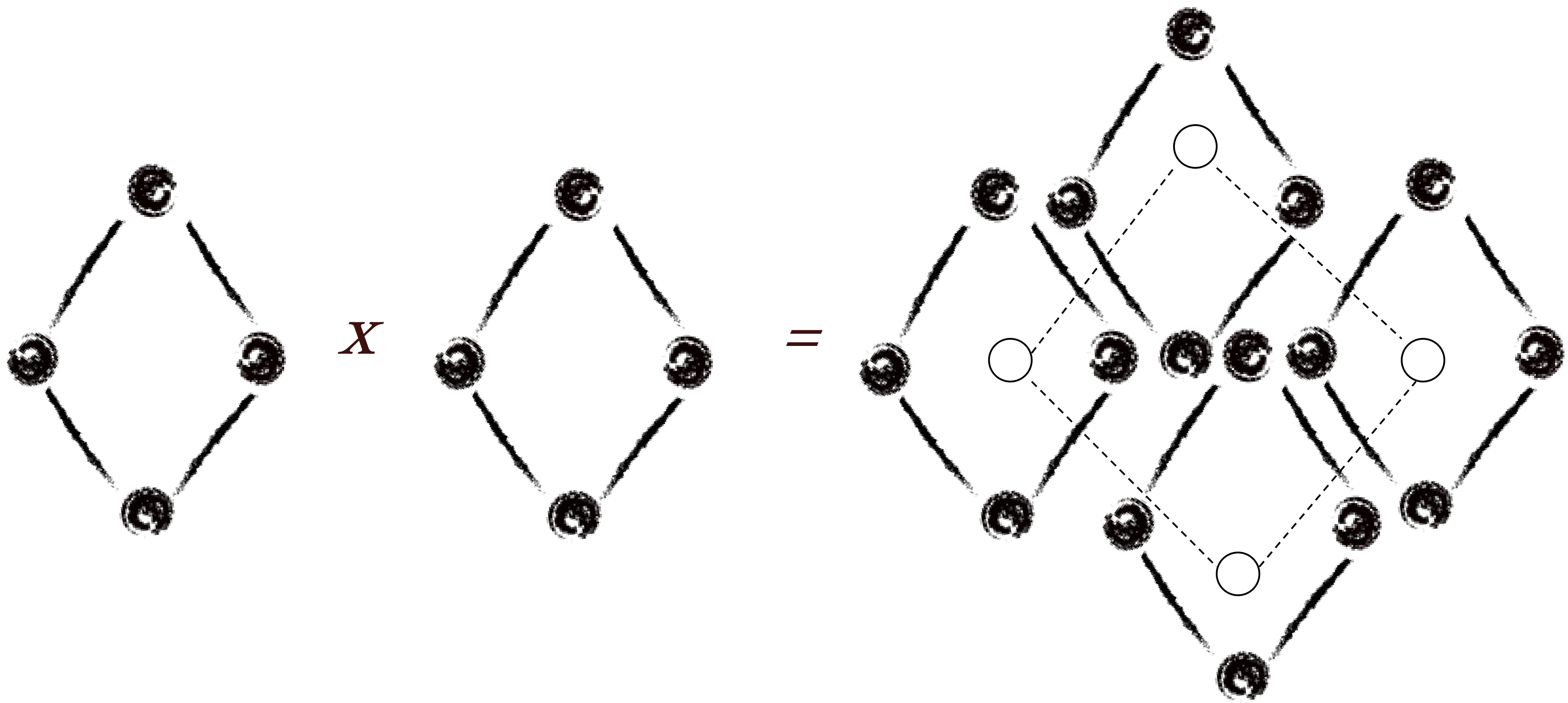
$$p(x \mid y \wedge i) = p(y \mid x \wedge i) \frac{p(x \mid i)}{p(y \mid i)}$$

$$\downarrow$$

$$p(x \mid y) = p(y \mid x) \frac{p(x \mid i)}{p(y \mid i)}$$

Bayes Theorem involves relating inferences under a change of context.

Lattice Products



Direct (Cartesian) product of two spaces

Direct Product Rule

The lattice product is also associative, commutative and closed

$$A \times (B \times C) \quad = \quad (A \times B) \times C$$

After the sum rule, the only freedom left is rescaling

$$p(a,b \mid i, j) \quad = \quad p(a \mid i)\, p(b \mid j)$$

which is again summation (under the invertible transform: logarithm)

Bayesian Probability Theory consists of Constraint Equations

Sum Rule

$$p(x \vee y \,|\, i) = p(x \,|\, i) + p(y \,|\, i) - p(x \wedge y \,|\, i)$$

Direct Product Rule

$$p(a, b \,|\, i, j) \quad = \quad p(a \,|\, i)\, p(b \,|\, j)$$

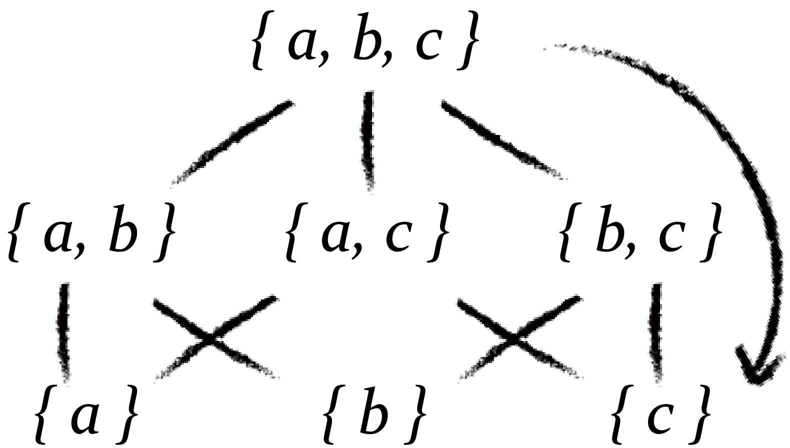
Product Rule

$$p(y \wedge z \,|\, x) = p(y \,|\, x)\, p(z \,|\, x \wedge y)$$

Bayes Theorem

$$p(x \,|\, y) = p(y \,|\, x) \frac{p(x \,|\, i)}{p(y \,|\, i)}$$

Inference



statements

Given a quantification of the join-irreducible elements, one uses the constraint equations to consistently assign any desired bi-valuations (probability)

This derivation gives meaning to *probability* as the degree of implication

How far can we take these ideas?

One can derive:

Information Theory

Feynman Path Integral Formulation of Quantum Mechanics

Special Relativity

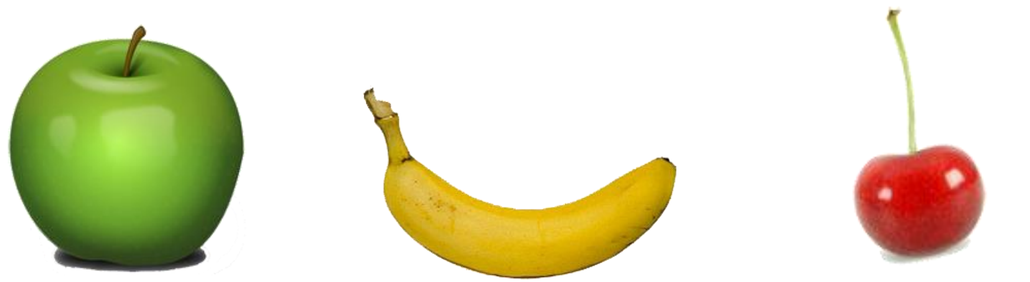
a b c

Choosing a Piece of Fruit

apple

banana

cherry



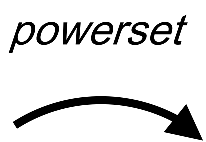
apple banana cherry

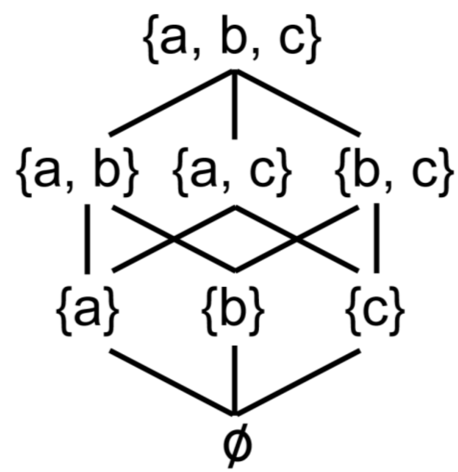
States describe Systems
Antichain

Potential States given by Powerset

$\overline{N}$ 2^{N}

a b c





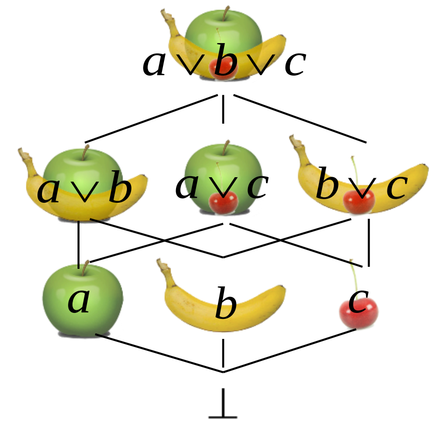
Potential States

$\overline{N}$

a **b** **c**

powerset

2^N



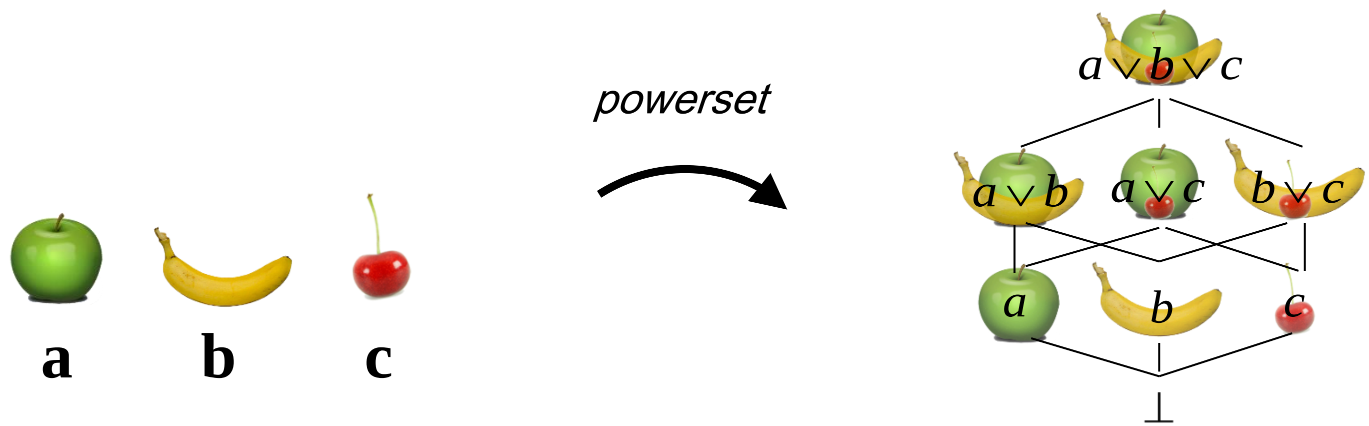
$$a \doteq \{\mathrm{a}\}$$

$$a \vee b \doteq \{\mathrm{a}, \mathrm{b}\}$$

$$\rightarrow \doteq \subseteq$$

Statements = Sets of Potential States

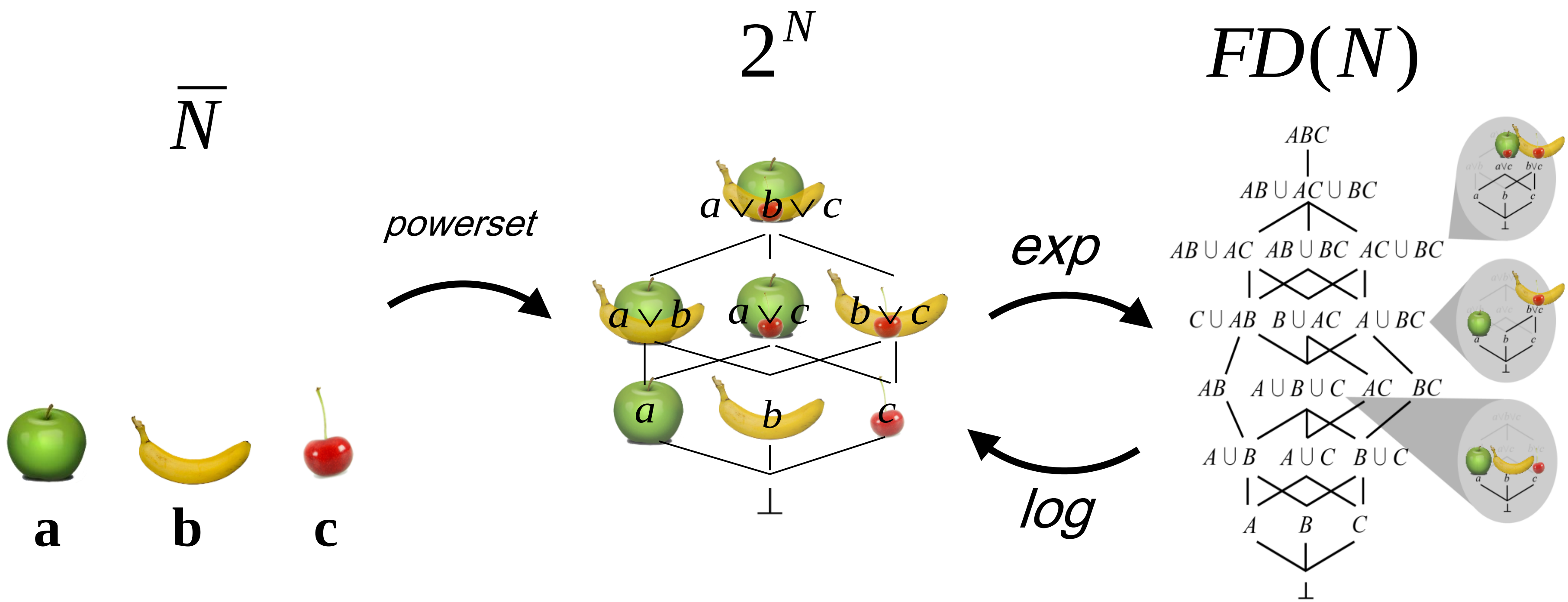
$\overline{N}$ 2^N



States

Statements
(sets of states)
(potential states)

Three Spaces



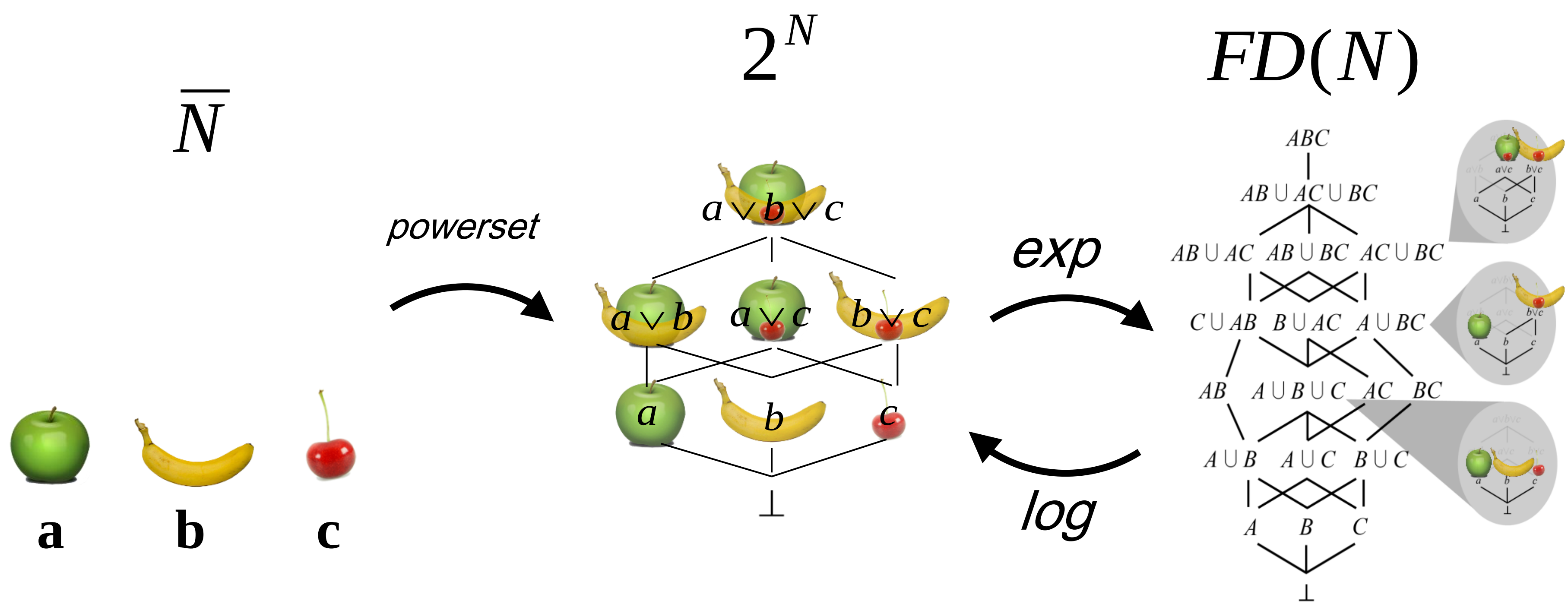
$$a \doteq \{\mathrm{a}\}$$

$$a \vee b \doteq \{\mathrm{a}, \mathrm{b}\}$$

$$A \doteq \{a\}$$

$$AB \doteq \{\mathrm{a}, \mathrm{b}, \mathrm{a} \vee b\}$$

Questions as Sets of Potential Statements



States

Statements
(sets of states)
(potential states)

Questions
(sets of statements)
(potential statements)

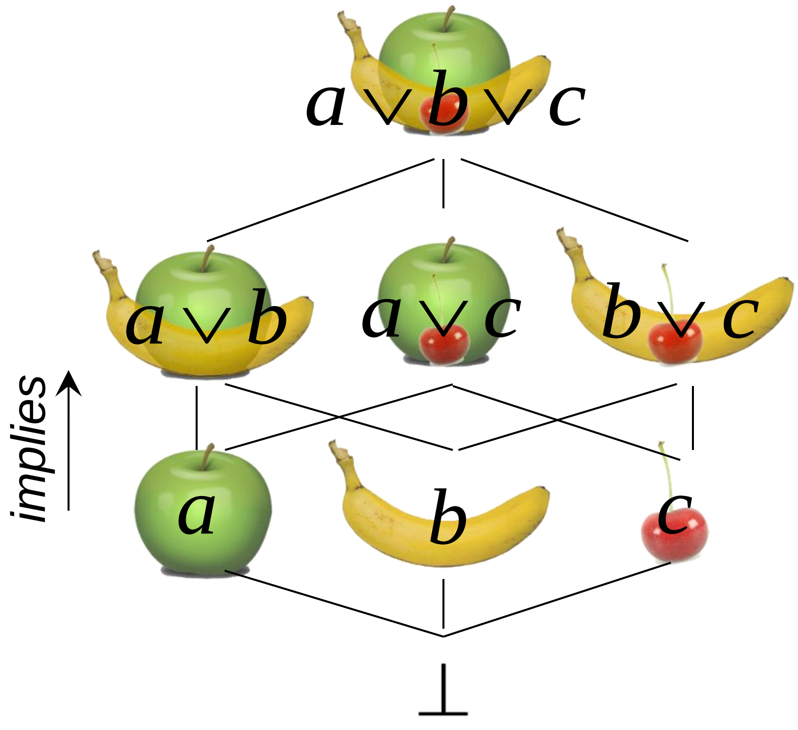
State Space

apple *banana* *cherry*

States describe Systems

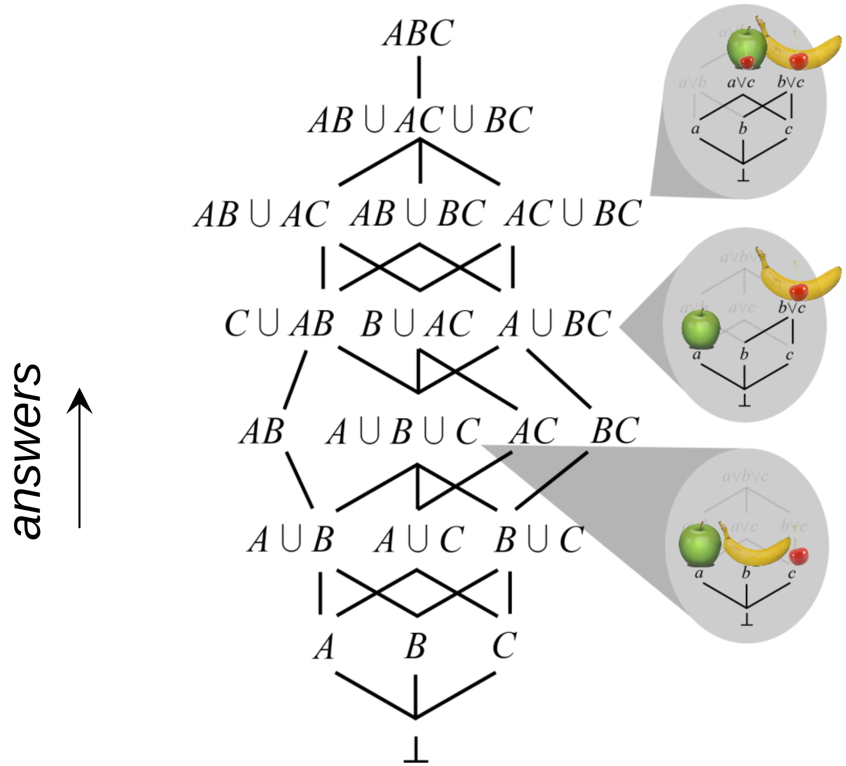
Antichain

Hypothesis Space (Space of Statements)



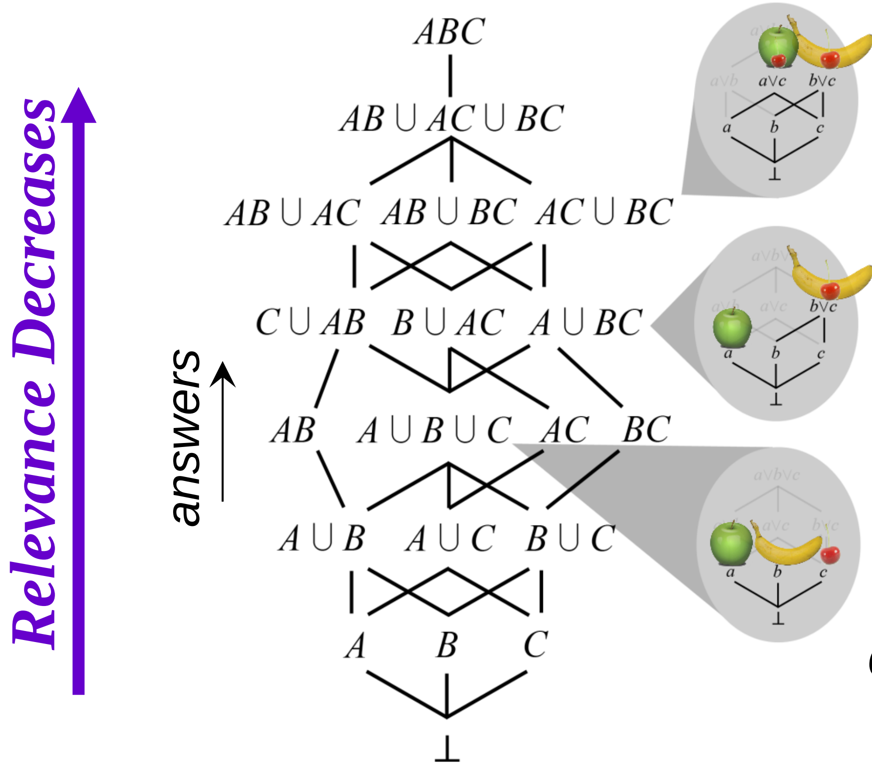
Statements are sets of Potential States
Boolean Lattice

Inquiry Space (Space of Questions)



Questions are sets (downsets) of Statements
Free Distributive Lattice

Questions Can Answer One Another



"Is it an Apple or Cherry, or is it a Banana or Cherry?"

"Is it an Apple?"

Central Issue
"Is it an Apple, Banana, or Cherry?"

Central Issue

I = "Is it an Apple, Banana, or Cherry?"

This question is answered by the following set of statements:

I = { a = "It is an Apple!",
b = "It is a Banana!",
c = "It is a Cherry!" }

$$I = \{a, b, c\}$$

Questions Can Answer One Another

Now consider the binary question

B = "Is it an Apple or not an Apple?"

B = {*a* = *"It is an Apple!"*, ~*a* = *"It is not an Apple!"*}

$$B = \{a, b \vee c, b, c\}$$

As the defining set is exhaustive, $\sim a = b \vee c$

Ordering Questions and Answering

I = "Is it an Apple, Banana, or Cherry?"

$$I = \{a, b, c\}$$

B = "Is it an Apple?"

$$B = \{a, b \vee c, b, c\}$$

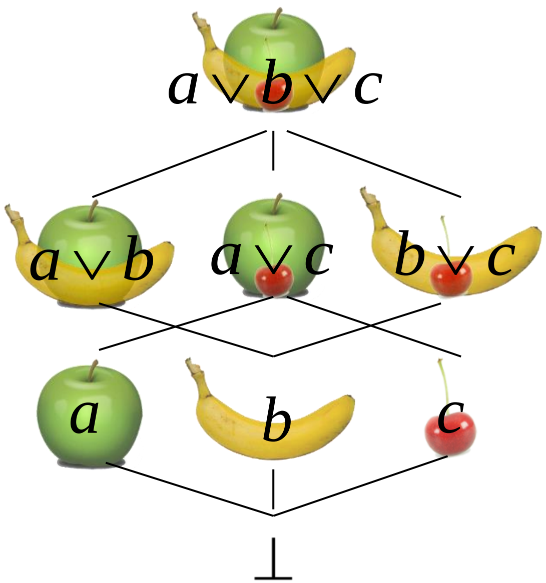
$$I \subseteq B$$

$\Rightarrow$

I answers B

B includes I

Probability and Statements



Probability quantifies the degree to which one statement implies another

$$p(x \mid i)$$

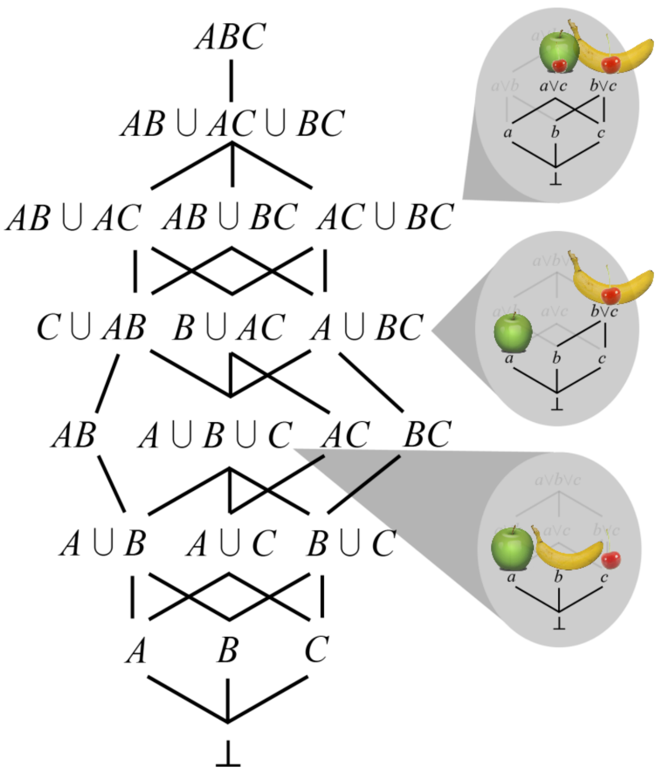
Constraint Equations

$$p(x \vee y \mid i) = p(x \mid i) + p(y \mid i) - p(x \wedge y \mid i)$$

$$p(x \wedge y \mid i) = p(x \mid i)\; p(y \mid x \wedge i)$$

$$p(x \mid y \wedge t) = \frac{p(x \mid t)\, p(y \mid x \wedge t)}{p(y \mid t)}$$

Relevance and Questions



Relevance quantifies the degree to which one question answers another

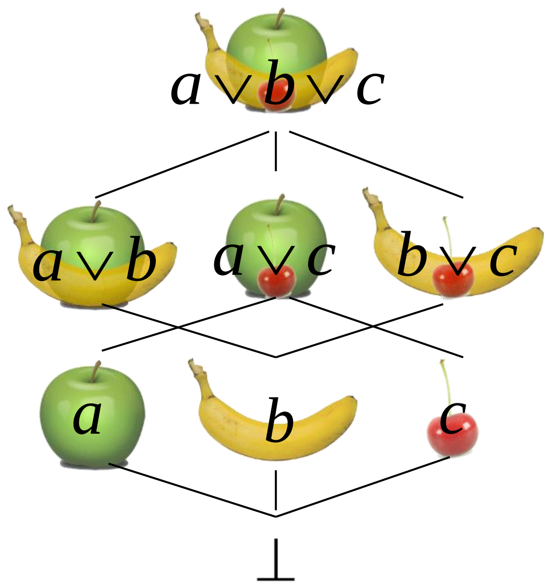
$$d(X \mid Y)$$

Constraint Equations

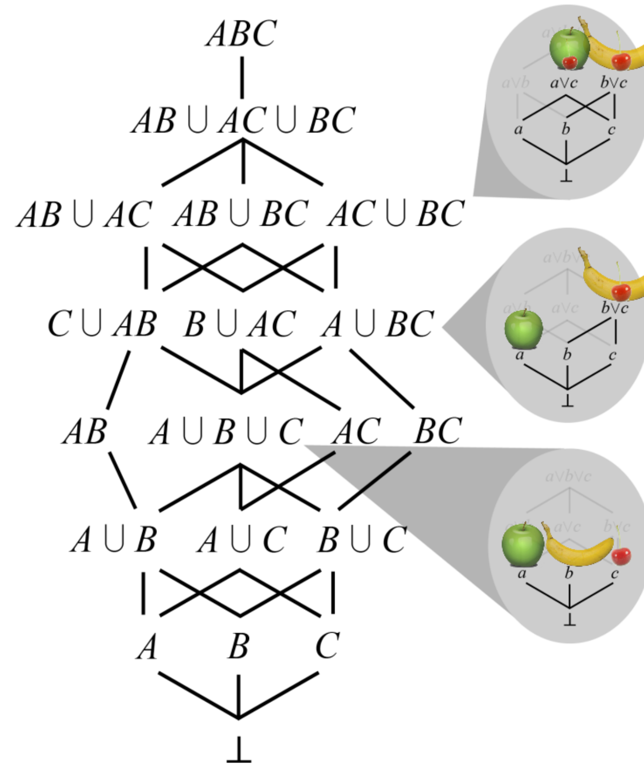
$$d(X \vee Y \mid Z) = d(X \mid Z) + d(Y \mid Z) - d(X \wedge Y \mid Z)$$

$$d(X \mid Z) = d(X \mid Y)\, d(Y \mid Z)$$

Probability and Relevance

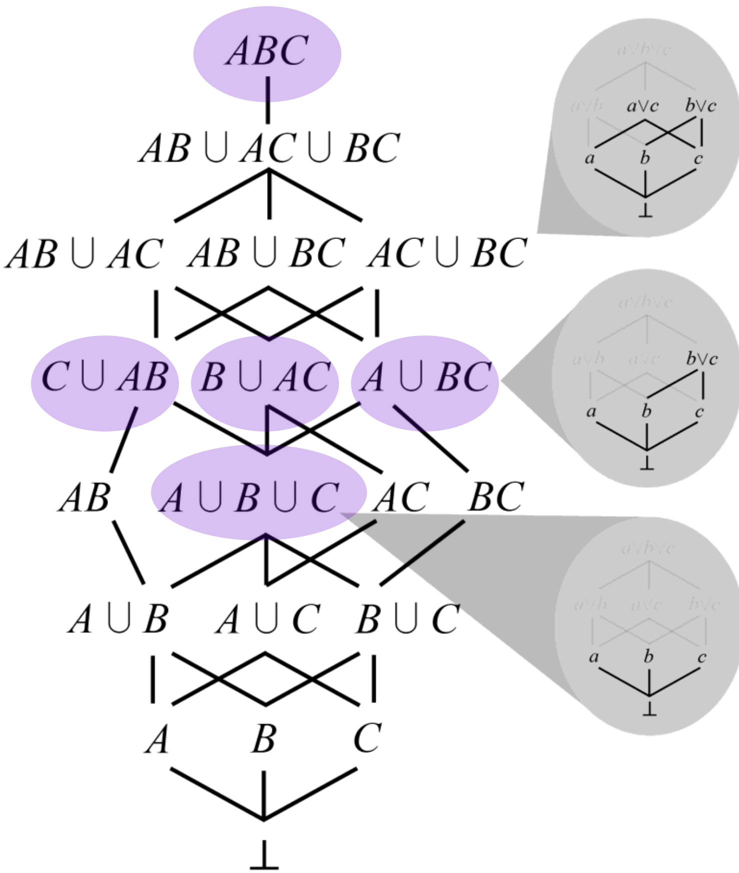


Relevance is a function of probability

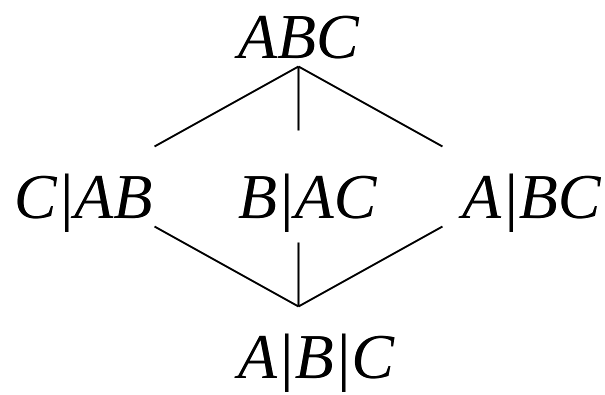


FURTHER ASSERT *that the degree to which one question answers another must depend on the probabilities of the possible answers.*

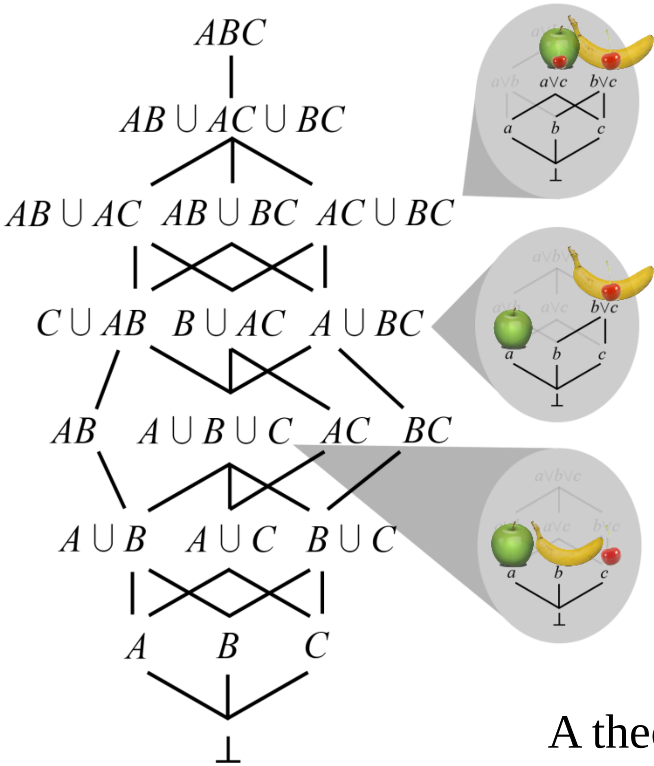
Partition Questions



One can show that relevance is only a valid measure on the sublattice of questions isomorphic to partitions



Relevance and Entropy



$$d(I \mid Q) = aH(Q) + b$$

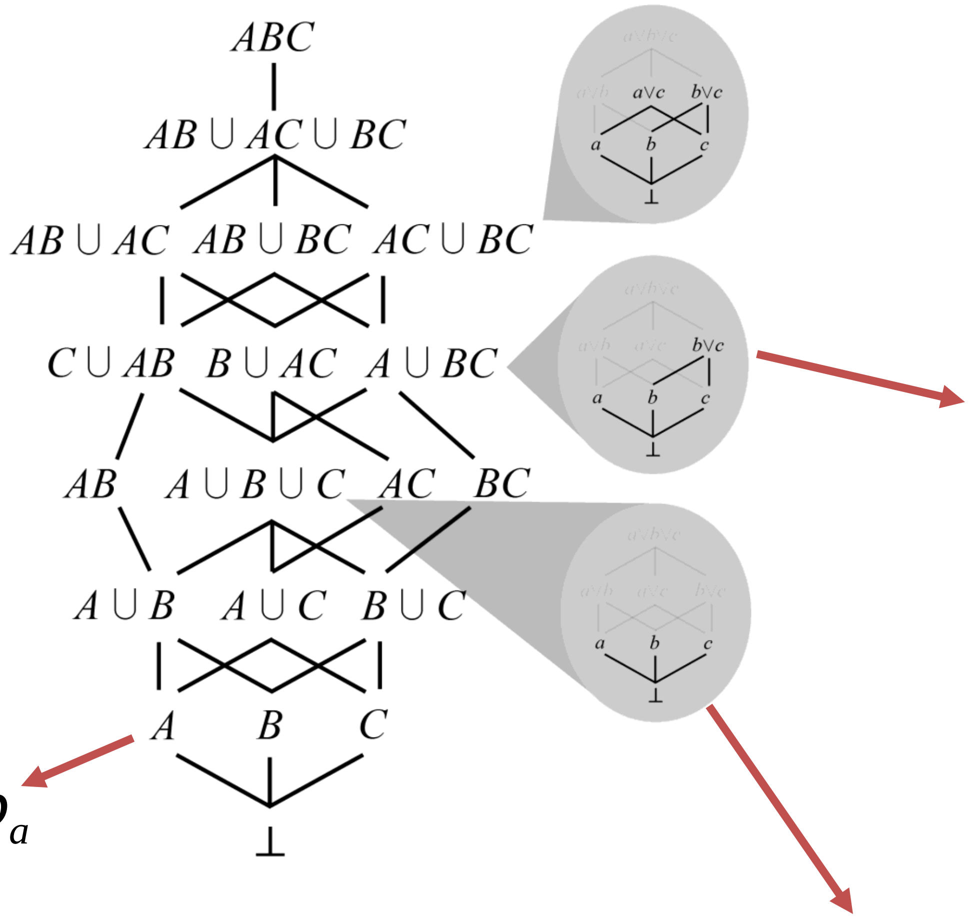
$$= -a \sum_{i=1}^{n} p_i \log_2 p_i + b$$

A theorem by Aczel and Ng further constrains the relevance, such that **the degree to which a partition question answers the central issue is proportional to the Shannon entropy of the partition questions top answers**.

Relevance and Entropy

One can normalize with respect to $H(I)$

$$d(I\,|\,Q)$$



$$H(p_a, p_{b\vee c})$$

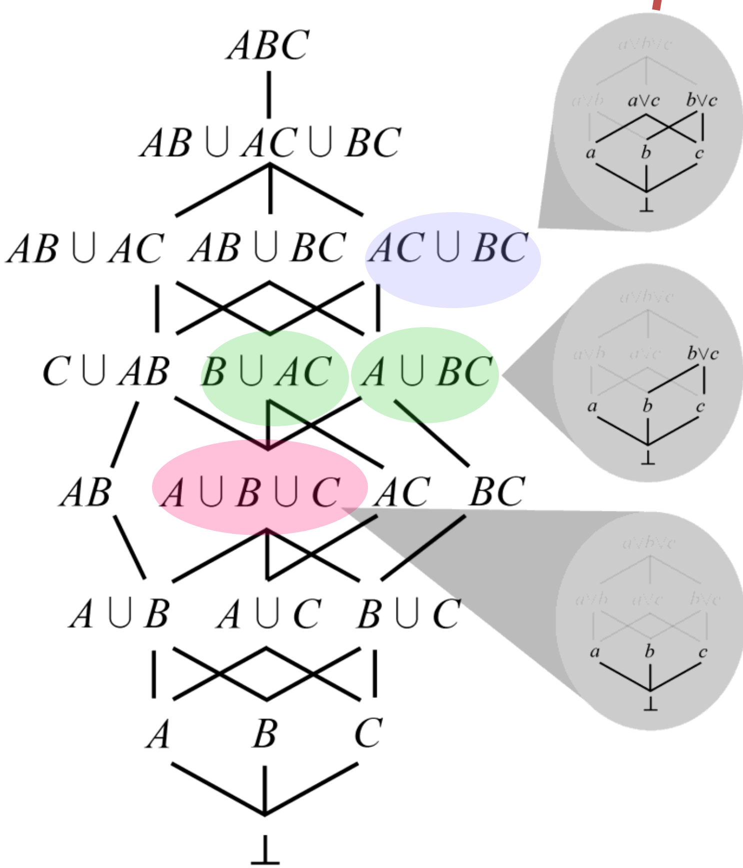
$$-\,p_a \log_2 p_a$$

$$H(I) = -\,p_a \log_2 p_a - p_b \log_2 p_b - p_c \log_2 p_c$$

Higher Order Informations

$$d(AC \cup BC \mid I) = d(B \cup AC \mid I) + d(A \cup BC \mid I) - d((B \cup AC) \wedge (A \cup BC) \mid I)$$

$$d(I \mid AC \cup BC) \sim I(B \cup AC; A \cup BC)$$



This relevance is related to the mutual information.

In this way one can obtain higher-order informations.

However, often these are invalid as they may involve non-partition questions.

Guessing Game

apple *banana* *cherry*

Can only ask binary (YES or NO) questions!

Which Question to Ask?

Is it or is it not an Apple?

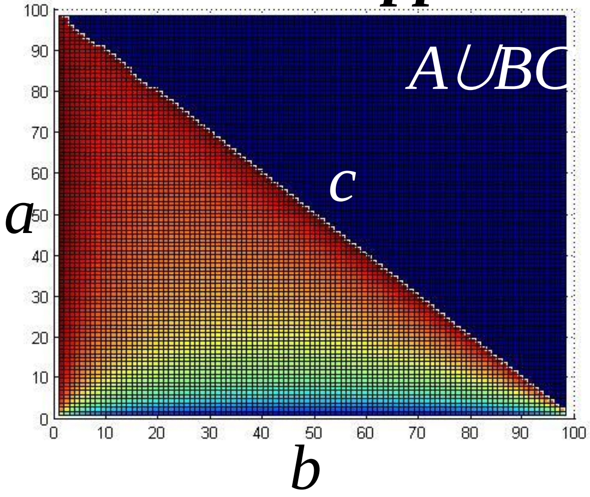
Is it or is it not a Banana?

Is it or is it not a Cherry?

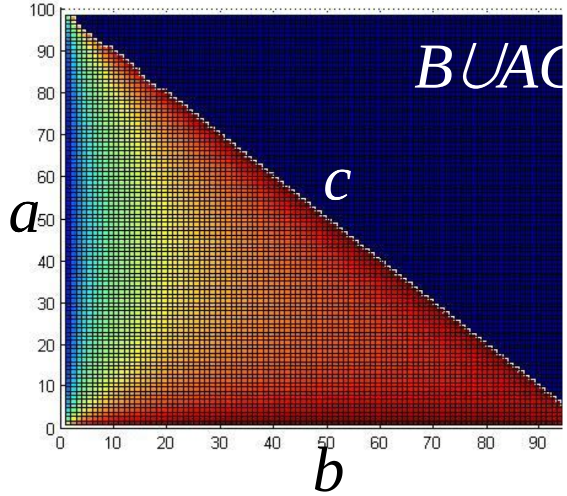
If you believe that there is a
75% chance that it is an Apple,
and a 10% chance that it is a Banana,
which question do you ask?

Relevance Depends on Probability

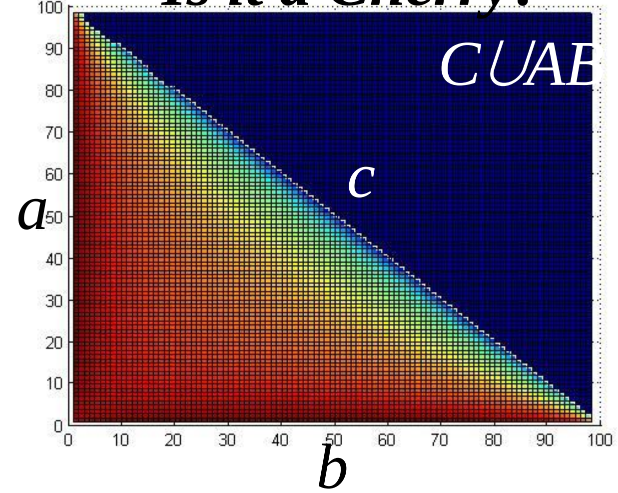
Is it an Apple?



Is it a Banana?

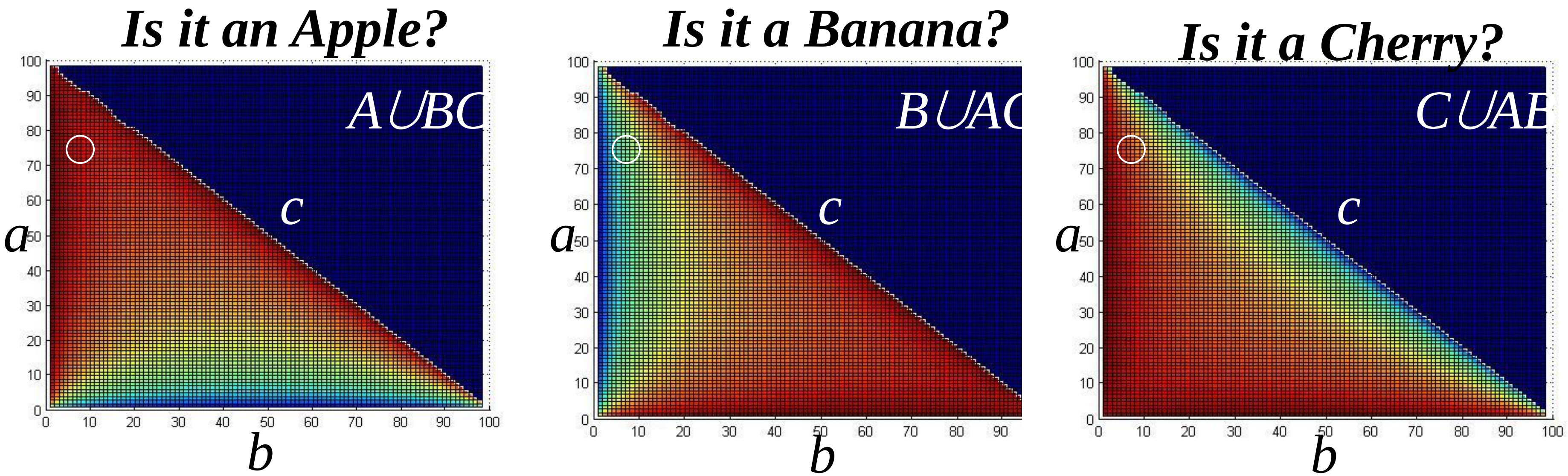


Is it a Cherry?



If you believe that there is a
75% chance that it is an Apple,
and a 10% chance that it is a Banana,
which question do you ask?

Relevance Depends on Probability



$d(I \mid A \cup BC) \propto 0.5623$ $d(I \mid B \cup AC) \propto 0.3250$ $d(I \mid C \cup AB) \propto 0.4227$

If you believe that there is a
75% chance that it is an Apple,
and a 10% chance that it is a Banana,
which question do you ask?

Earth Science Research Team

Kevin H. Knuth, PI
Univ at Albany (SUNY)

Deniz Gençağa
Carnegie Mellon Univ

William B. Rossow
City College of New York
(formerly NASA GISS)

FUNDING: NASA ESTO
Advanced Information Systems Technology, Knuth (PI)
Cloud Modeling and Analysis Initiative, Rossow (PI), Knuth (co-I)

Lorenz System

$$\dot{x} = \sigma(y - x)$$

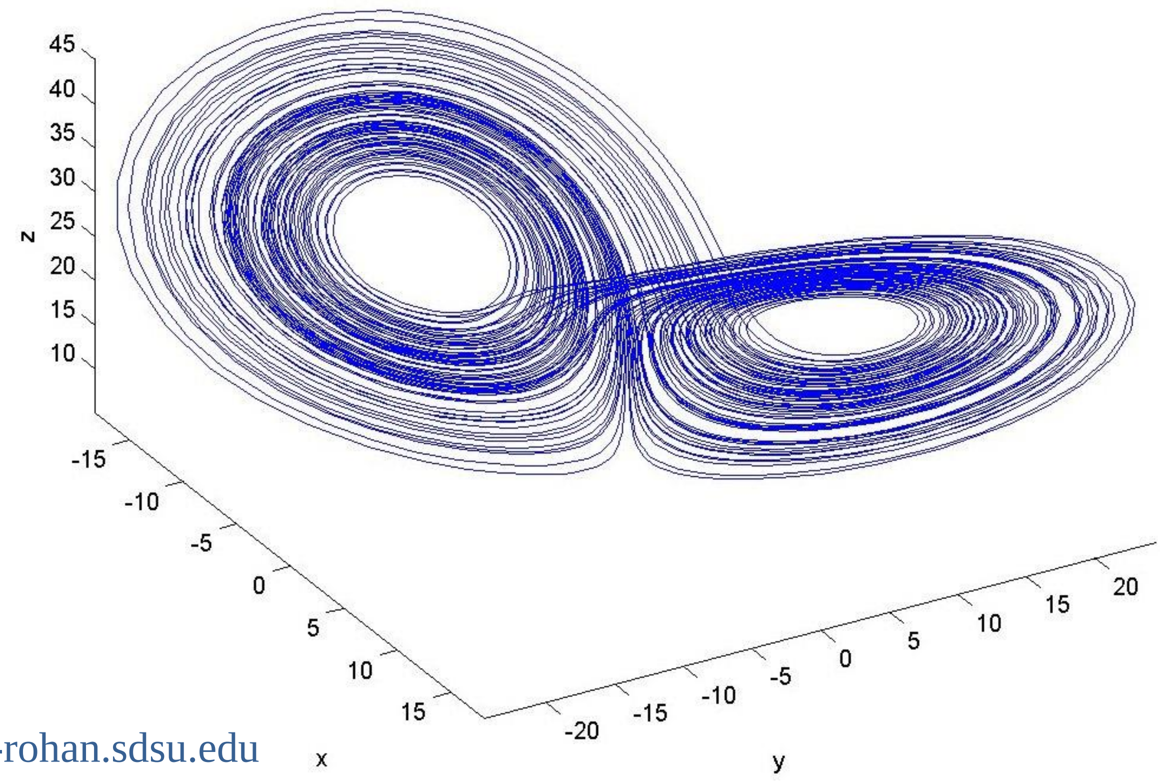
$$\dot{y} = -xz + rx - y$$

$$\dot{z} = xy - bz$$

$\sigma = 10$

$b = 8/3$

$r = Rayleigh\ Number$



www-rohan.sdsu.edu

Lorenz System

How do these variable influence one another?
NOT OBVIOUS!

$$\dot{x} = \sigma(y - x)$$

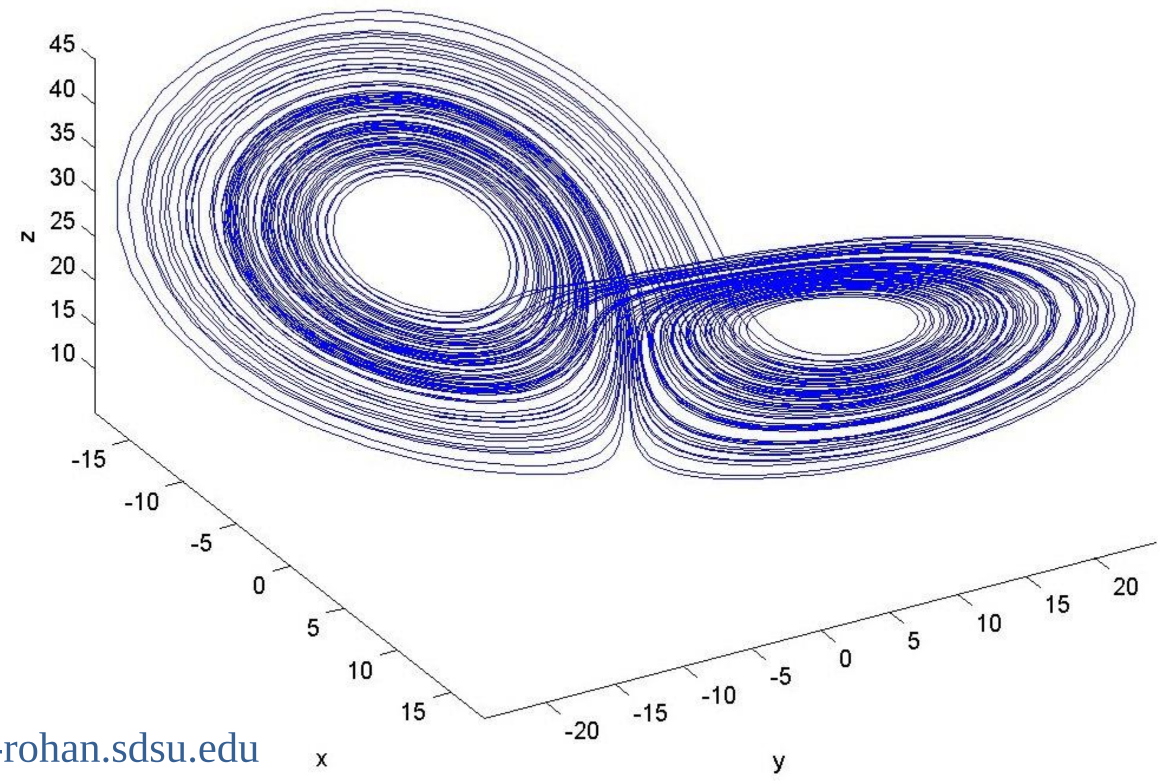
$$\dot{y} = -xz + rx - y$$

$$\dot{z} = xy - bz$$

$\sigma = 10$

$b = 8/3$

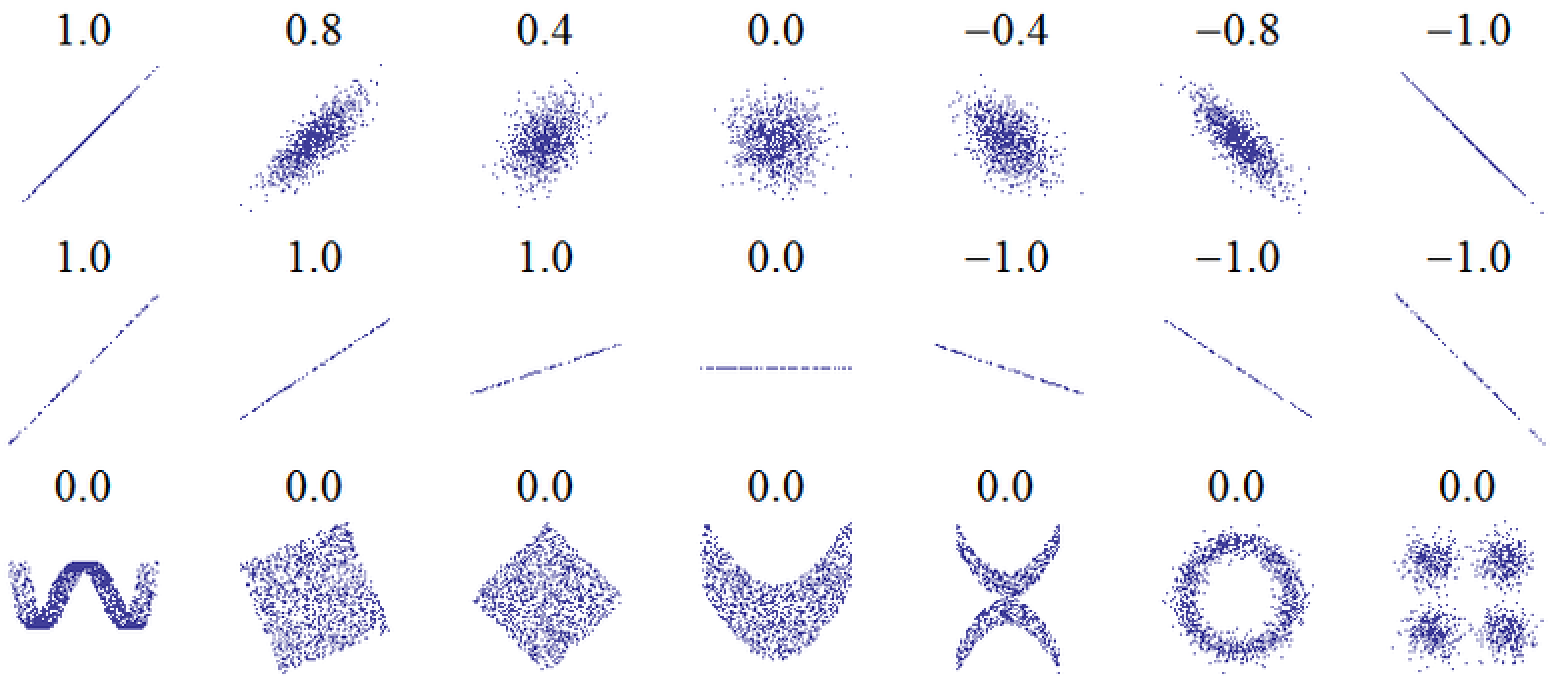
$r = Rayleigh\ Number$



www-rohan.sdsu.edu

Correlation Coefficient Examples

Joint Distributions of Two Variables *X* and *Y*



http://en.wikipedia.org/wiki/File:Correlation_examples.png

Decorrelation does not mean Independent

DE-CORRELATED ≠ INDEPENDENT

Entropy

We use x to denote the state of the system out of a set of possible states X

The **surprise** is large for improbable states and small for probable states.

$$h(x) = \log \frac{1}{p(x)}$$

Averaging this quantity over all of the possible states of the system gives a measure of our knowledge about the state of the system

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log p(x)$$

which is called the **entropy**.

Mutual Information

An important quantity is given by the sum and difference of entropies,

$$MI(X,Y) \quad = \quad H(X) + H(Y) - H(X,Y)$$

This is called the **Mutual Information** (MI) since it describes the amount of information that is shared between the two subsystems.

$$MI(X,Y) \quad = \quad \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

Mutual Information is zero if X and Y are statistically independent. However, it is never zero in practice when computed from data.

Need to quantify uncertainties!

Transfer Entropy

Schreiber (2000) introduced an information-theoretic quantity called the **Transfer Entropy** (TE). Consider two subsystems X and Y, with data in the form of a two time series of measurements

$$X = \{x_1, x_2, \cdots, x_t, x_{t+1}, \cdots, x_n\}$$
$$Y = \{y_1, y_2, \cdots, y_s, y_{s+1}, \cdots, y_n\}$$

then the transfer entropy can be written as

$$T(X_{t+1} \mid X_t, Y_s) \quad = \quad -H(X_t) + H(X_t, Y_s) + H(X_t, X_{t+1}) - H(X_t, X_{t+1}, Y_s)$$

which describes the degree to which information about Y allows one to predict future values of X. This is a potential measure of the causal influence that the subsystem Y has on the subsystem X.

The concepts behind the procedure are straightforward:

1. **Estimate the probability density from which the data were sampled.**
2. **Using this probability density, estimate the various necessary entropies.**

Challenges

First, difficult to perform objectively since probability density models often have free parameters that must be assigned.

Second, we interested in the values of these quantities, but we are also interested in the associated uncertainties of our estimates.

Third, Even worse, the entropy of the most probable density model does not correspond to the most probable entropy!
(Jacobians come in to play)

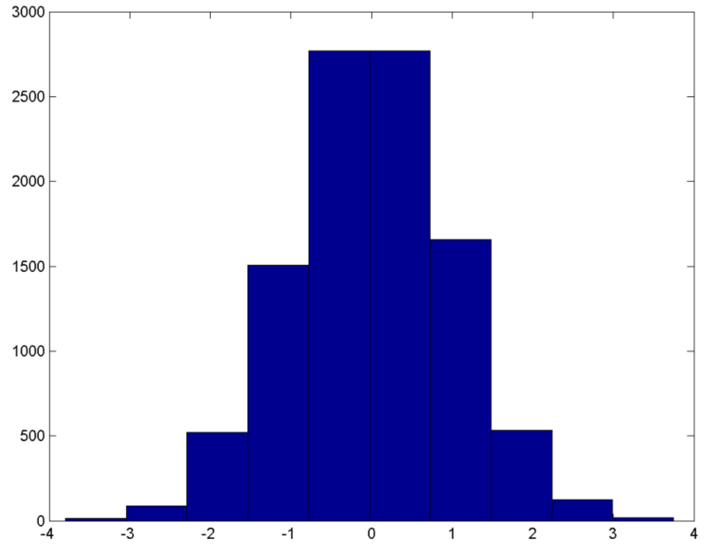
Challenges

First, difficult to perform objectively since probability density models often have free parameters that must be assigned.

Second, we interested in the values of these quantities, but we are also interested in the associated uncertainties of our estimates.

Third, Even worse, the entropy of the most probable density model does not correspond to the most probable entropy!
(Jacobians come in to play)

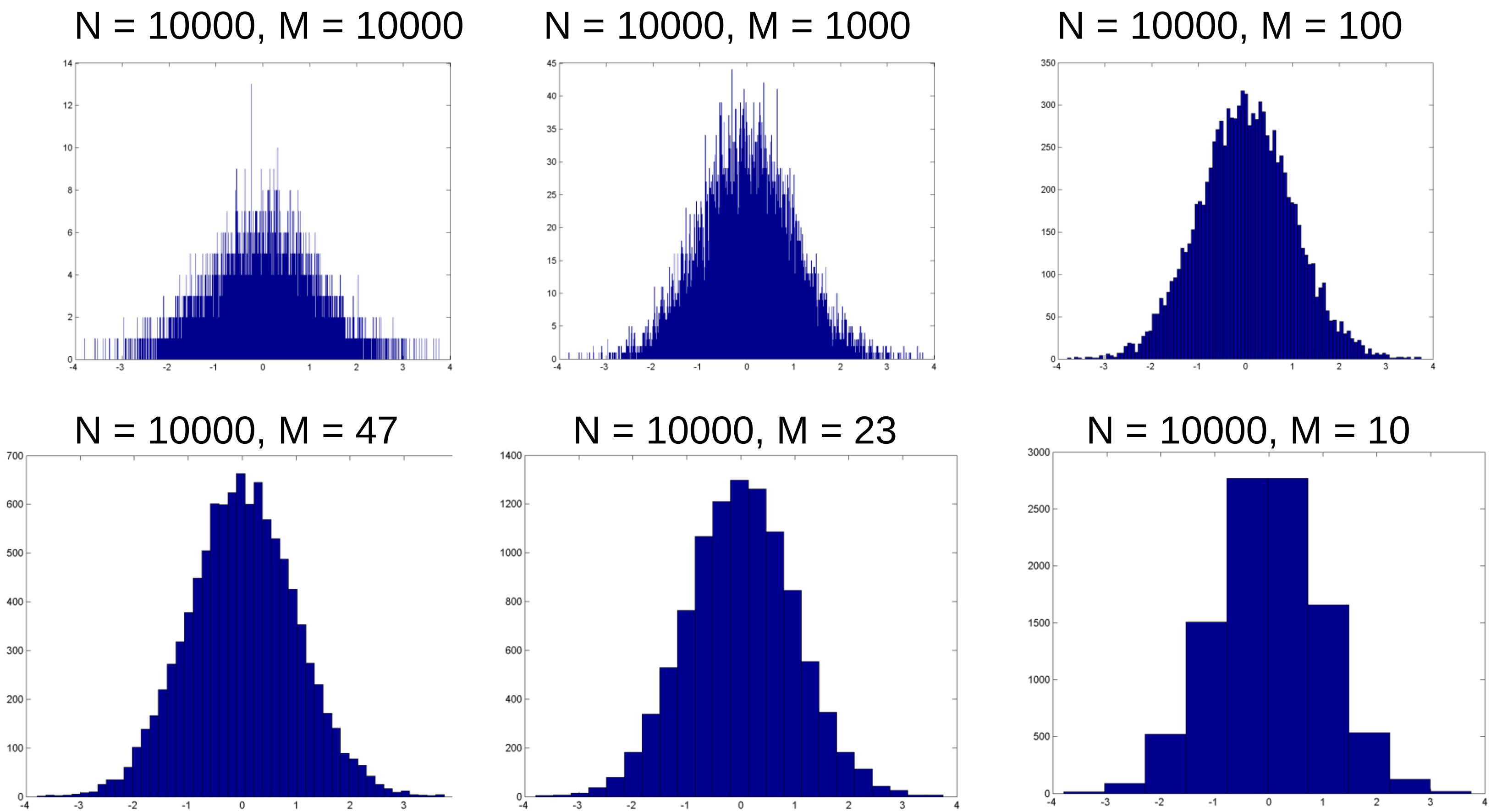
Histograms as Probability Density Models



Histograms can be viewed as simple models of the probability density from which the data were sampled.

They are convenient since they have regions of constant probability.

Histograms



The histogram should contain only details warranted by the data.
But how do we choose the Number of Bins?

Bayesian Posterior for the Number of Bins

By integrating over all possible bin probabilities, we can derive the posterior probability of the number of bins given the data.

$$p(M \mid \mathbf{d}, I) \;\propto\; \left(\frac{M}{V}\right)^{N} \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \frac{\prod_{k=1}^{M} \Gamma\left(n_k + \frac{1}{2}\right)}{\Gamma\left(n_1 + b_1 + \frac{3}{2}\right)}$$

It is easier to **find the number of bins that maximizes the logarithm of the posterior probability**

$$\log p(M \mid \mathbf{d}, I) \;=$$

$$N \log M \;+\; \log \Gamma\left(\frac{M}{2}\right) \;-\; M \log \Gamma\left(\frac{1}{2}\right) \;-\; \log \Gamma\left(N + \frac{M}{2}\right) \;+\; \sum_{k=1}^{M} \log \Gamma\left(n_k + \frac{1}{2}\right) + K$$

where K is the implicit proportionality constant.

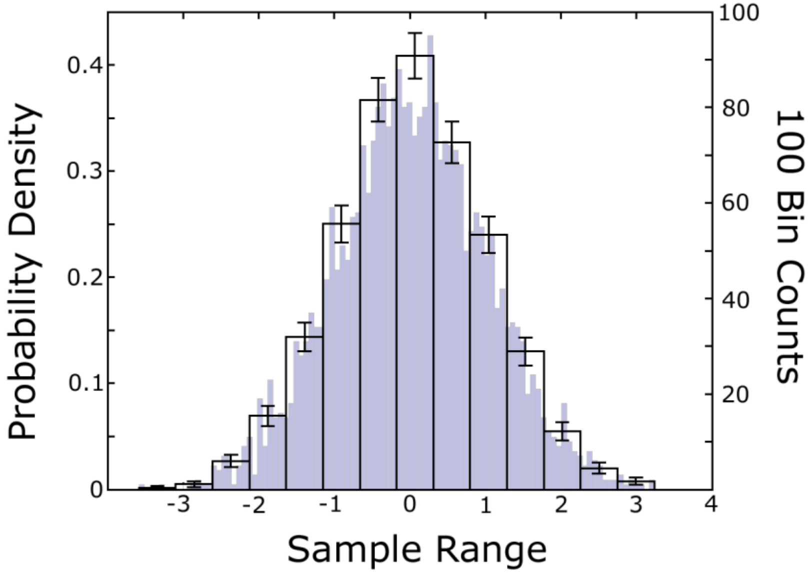
optBins Algorithm

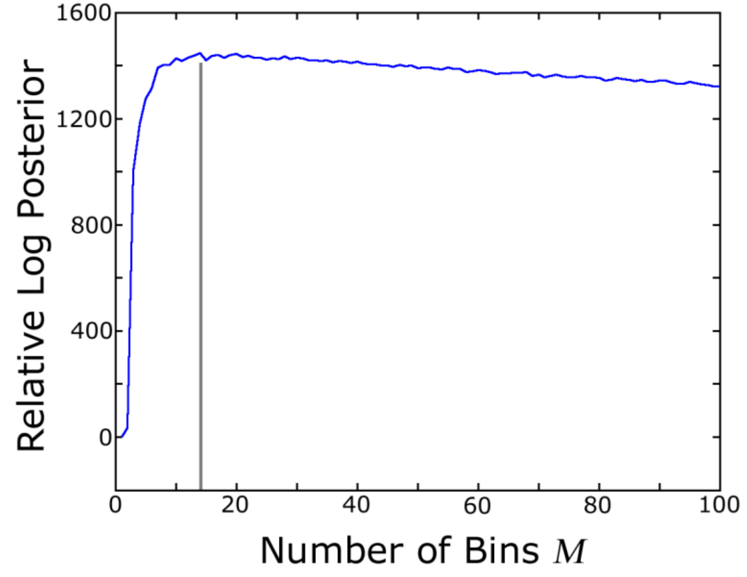
Now featured in Mathematica as the Knuth Method

```
function optM = optBINS(data,minM,maxM)
if size(data)>2 | size(data,1)>1
        error('data dimensions must be (1,N)');
    end
N = size(data,2);

% Loop through the different numbers of bins
% and compute the posterior probability for each.
logp = zeros(1,maxM);
for M = minM:maxM
    n = hist(data,M);  % Bin the data (equal width bins here)
    p = 0;
        for k = 1:M
          p = p + gammaln(n(k)+0.5);
        end
    logp(M) = N*log(M) + gammaln(M/2) - M*gammaln(1/2) - gammaln(N+M/2) + p;
end
[maximum, optM] = max(logp);
return
```

“Optimal” Histograms

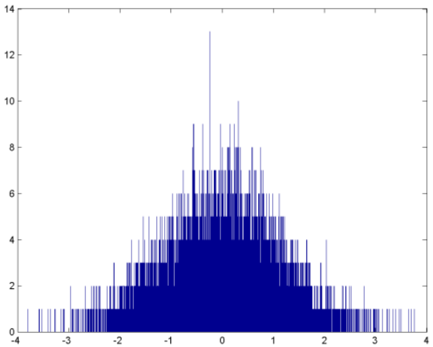




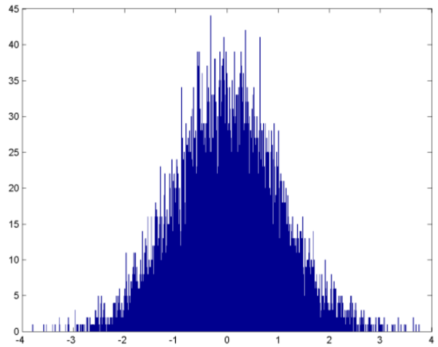
“Optimal” Binning for N = 3000 Gaussian distributed data points: **M = 14**

The "Optimal" Histogram

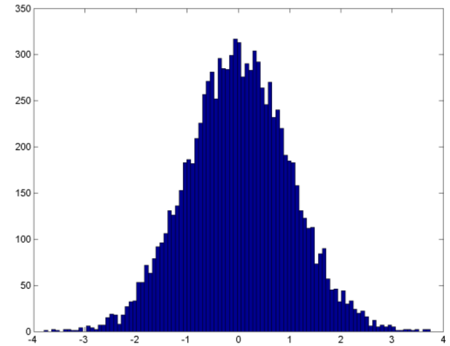
N = 10000, M = 10000



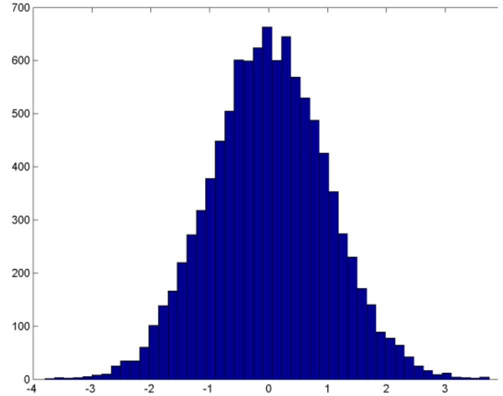
N = 10000, M = 1000



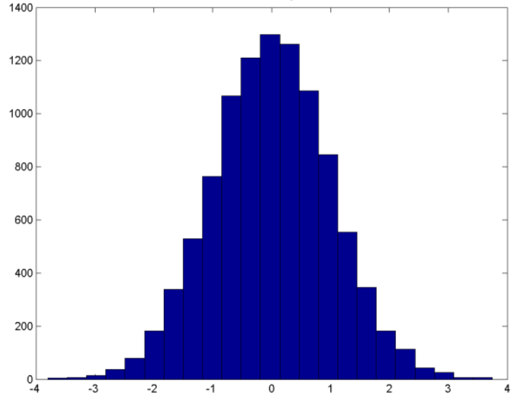
N = 10000, M = 100



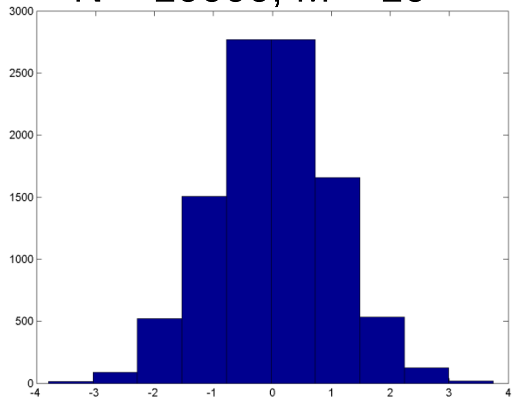
N = 10000, M = 47



N = 10000, M = 23



N = 10000, M = 10



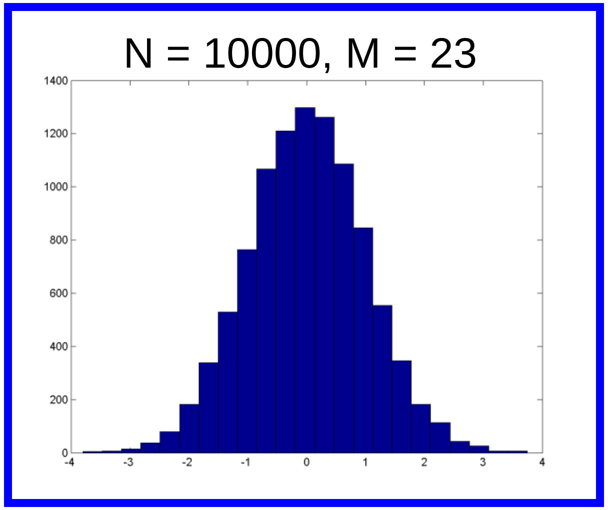
The histogram should contain only details warranted by the data.

Entropy Estimation

Entropy estimation is relatively easy with a constant-piecewise model

$$H = -\sum_i p_i \log p_i$$

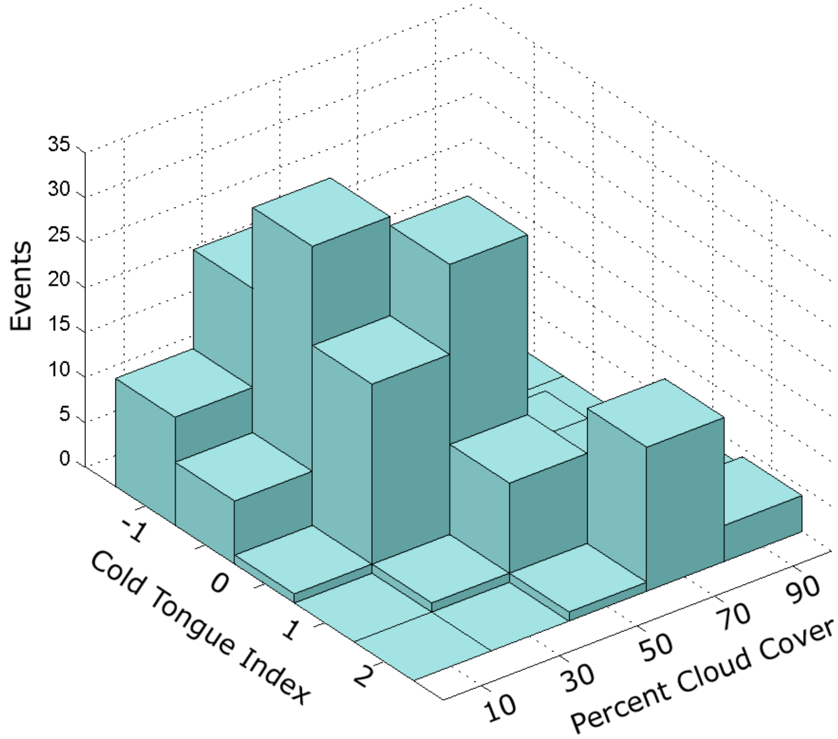
```
H = -sum(p .* (log(p) - log(vol)));
```



Entropy Estimation

And also in higher-dimensions…

$$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \log p(x,y)$$



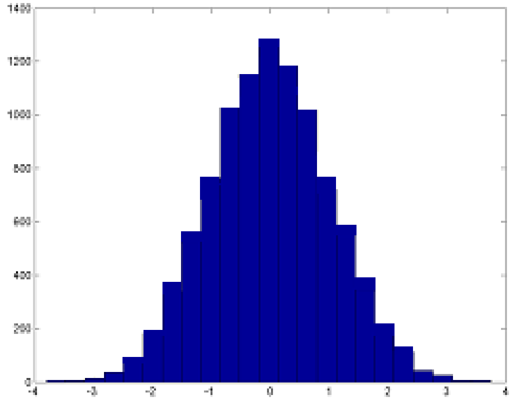
Estimating Uncertainties

To calculate the uncertainties in the entropy estimates, one must first realize that we are uncertain as to the bin probabilities of the probability density model.

By sampling a set of bin probabilities, we obtain a set of probable density functions, along with a set of probable entropies.

$$p(\boldsymbol{\pi}, M \mid \mathbf{d}, I) \propto \left(\frac{M}{V}\right)^{N} \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \pi_1^{n_1-\frac{1}{2}} \pi_2^{n_2-\frac{1}{2}} \ldots \pi_{M-1}^{n_{M-1}-\frac{1}{2}} \left(1-\sum_{k=1}^{M-1}\pi_k\right)^{n_M-\frac{1}{2}}$$

From this set of probable entropies, we can compute the mean and variance. Thus quantifying both the entropy and our uncertainty.



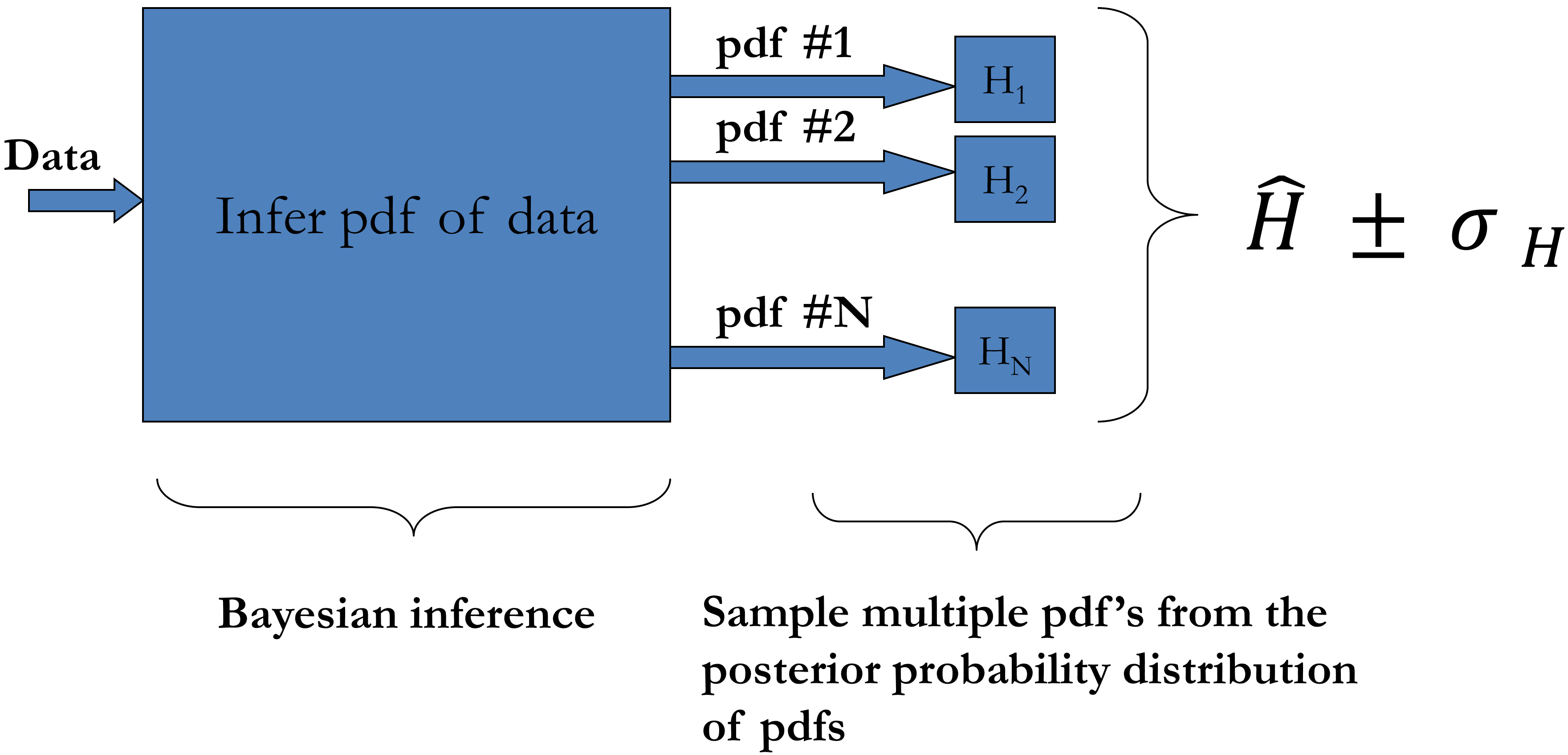
Challenges

First, difficult to perform objectively since probability density models often have free parameters that must be assigned.

Second, we interested in the values of these quantities, but we are also interested in the associated uncertainties of our estimates.

Third, Even worse, the entropy of the most probable density model does not correspond to the most probable entropy!
(Jacobians come in to play)

Estimating Entropy from Data



Estimating Information-Theoretic Quantities

Challenges

First, difficult to perform objectively since probability density models often have free parameters that must be assigned.

Second, we interested in the values of these quantities, but we are also interested in the associated uncertainties of our estimates.

Third, Even worse, the entropy of the most probable density model does not correspond to the most probable entropy!
(Jacobians come in to play)

Entropies from Sampling

This shows some of the results from sampling from the posterior probability and computing the entropies.

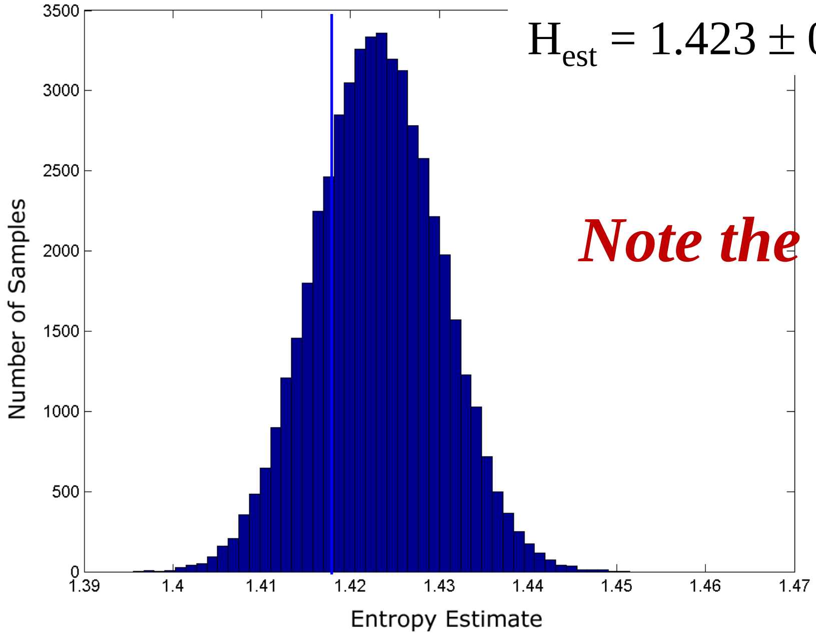
The data was from a Gaussian distribution with $\mu = 0$, $\sigma = 1$.

The true entropy is $H_{true} = 1.419$

$N = 10000$, $M = 24$

50000 Samples

H = 1.4202
1.4161
1.4159
…
1.4211
1.4259
1.4290



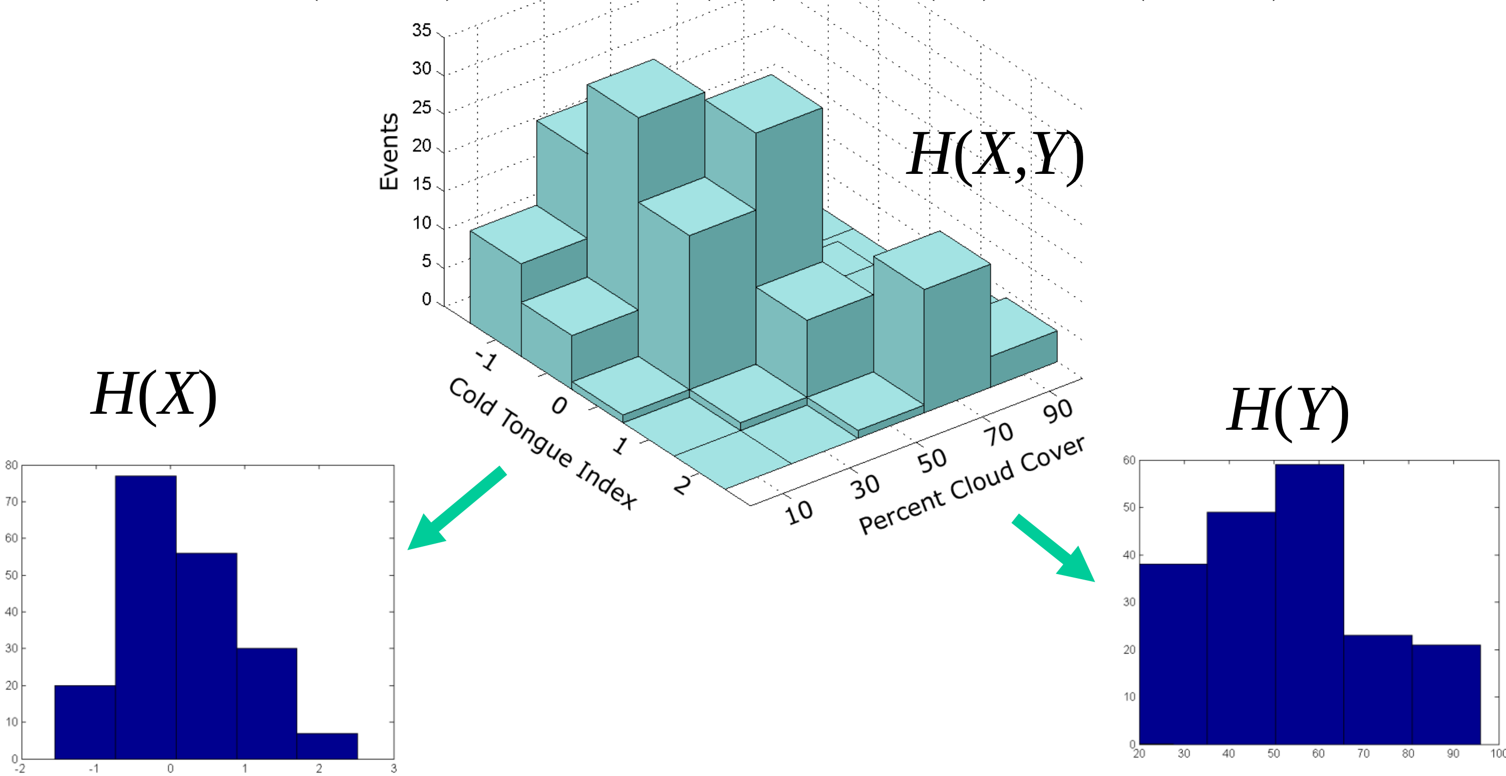
$H_{est} = 1.423 \pm 0.007$

Note the unavoidable bias

Estimating Mutual Information

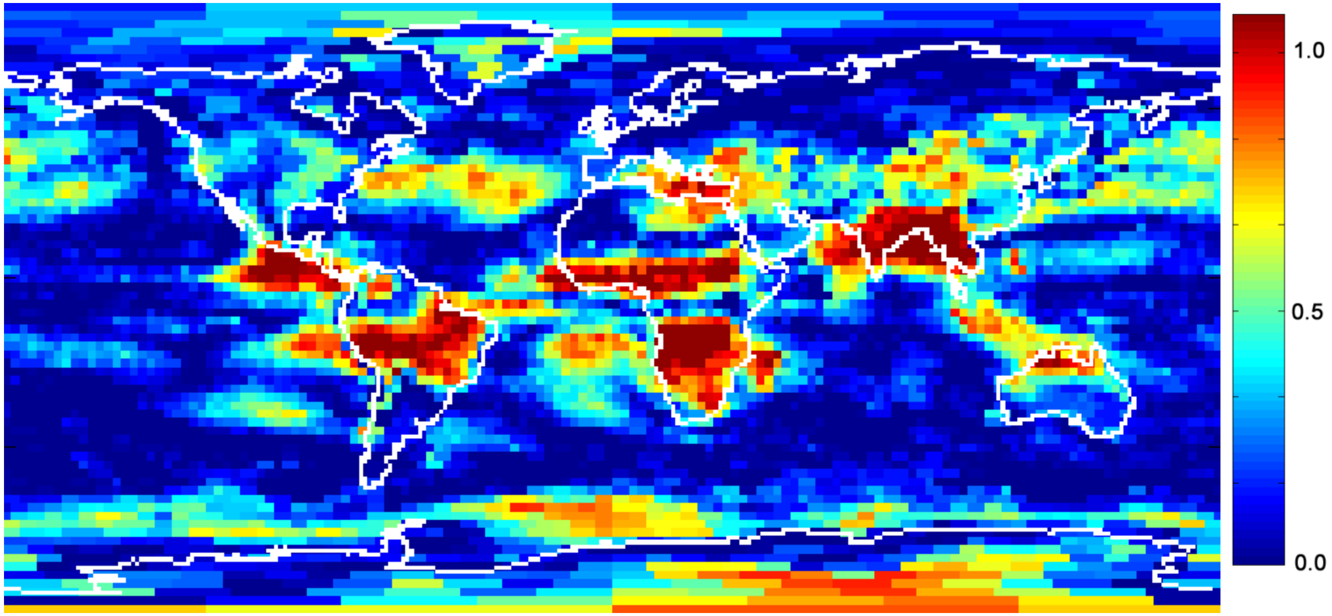
Mutual information requires the estimation of BOTH the two one-dimensional marginal entropies and two-dimensional joint entropy. We can use the same sampling strategy for all cases.

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$



Cloud Cover and Seasonality

Mutual Information between ISCCP percent cloud cover and Seasonality.

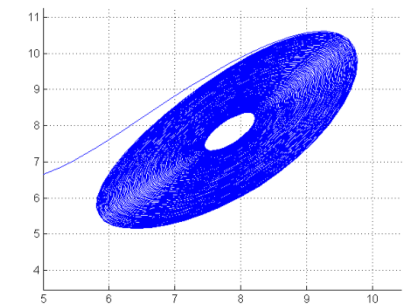


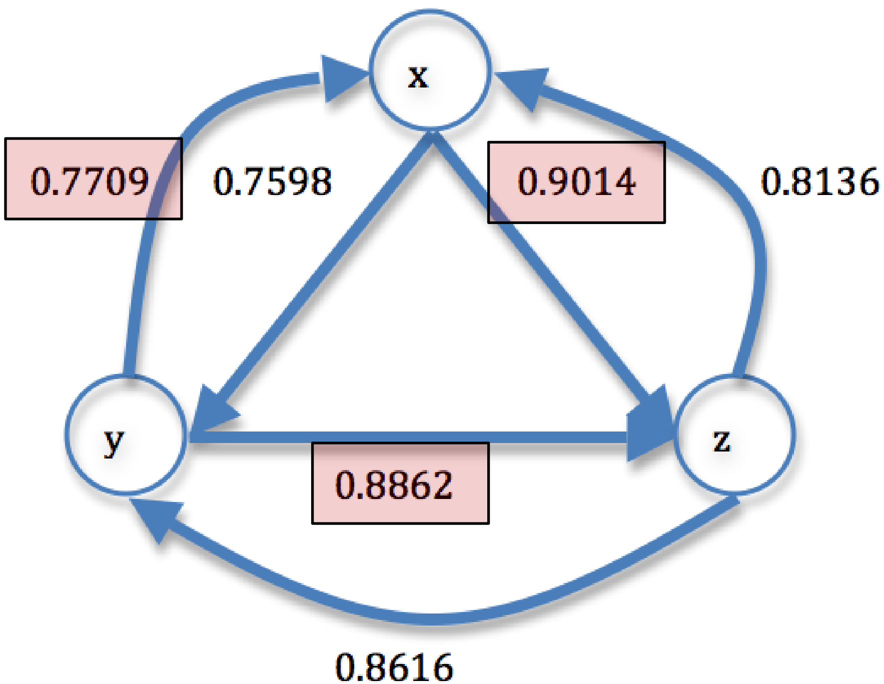
The data consisted of monthly averages of percent cloud cover resulting in a time-series of 198 months of 6596 equal-area pixels each with side length of 280 km.

This method finds the Inter-Tropical Convection Zones, The Monsoon Regions, the Sea Ice off Antarctica, and cloud cover in the North Atlantic and Pacific.

Transfer Entropy Results

Lorenz system r=24 (sub-chaotic regime)

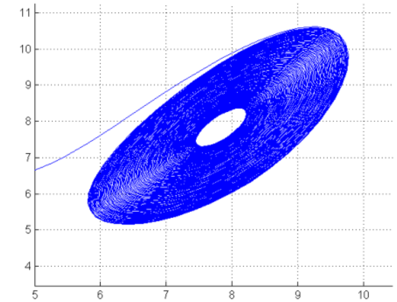


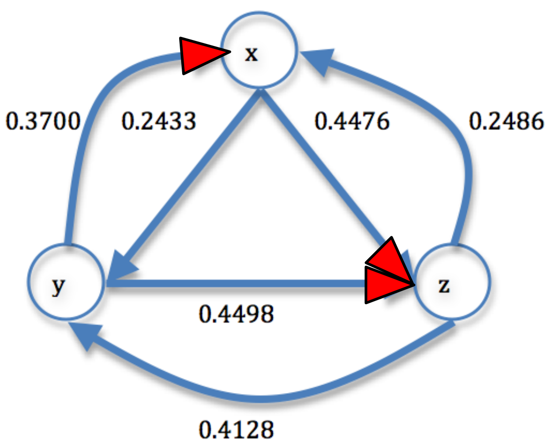


OptBINS Histogram Method $\beta = 0.1$

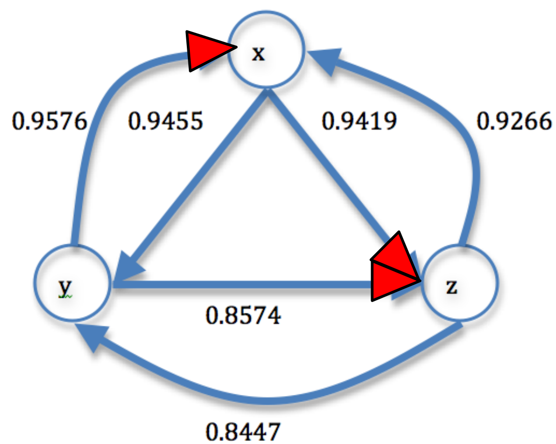
Transfer Entropy Results

Lorenz system r=24 (sub-chaotic regime)

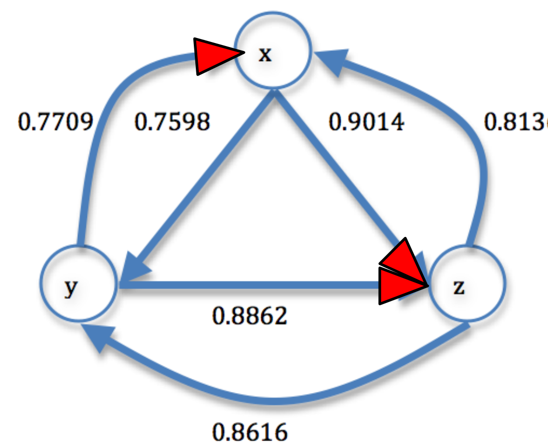




Kernel Density Estimation
(Prichard and Theiler, Grassberger & Procaccia)



Adaptive Partitioning
(Fraser & Swinney, Darbellay & Vajda)

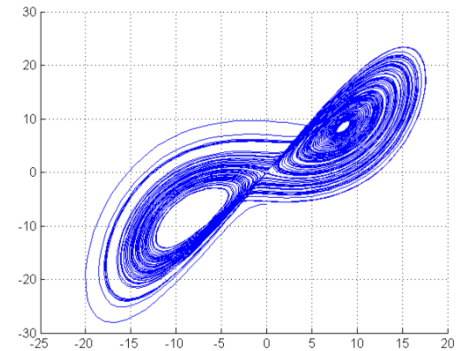


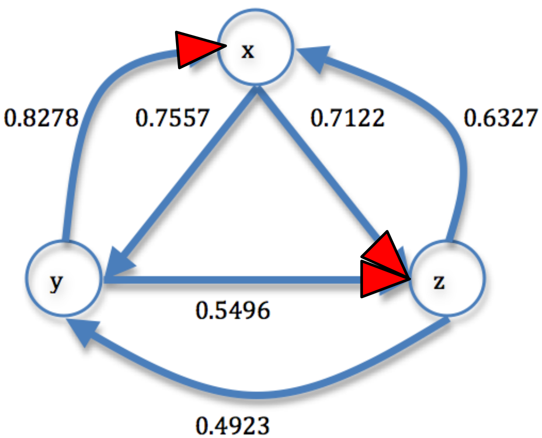
$\beta = 0.1$

OptBINS Histogram Method

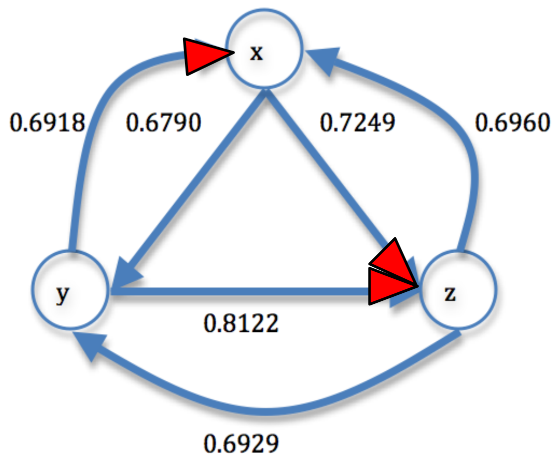
Transfer Entropy Results

Lorenz system r=28 (chaotic regime)

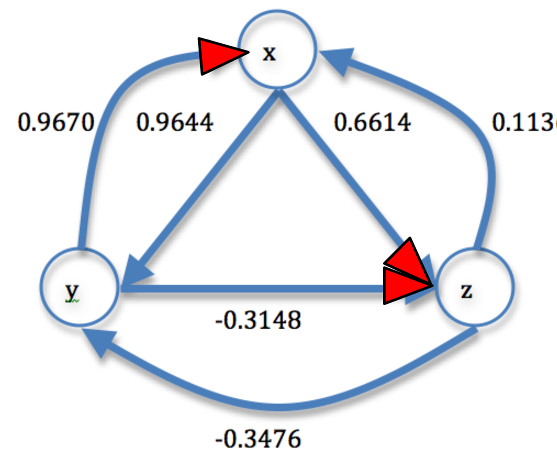




Kernel Density Estimation



Adaptive Partitioning



OptBINS Histogram Method

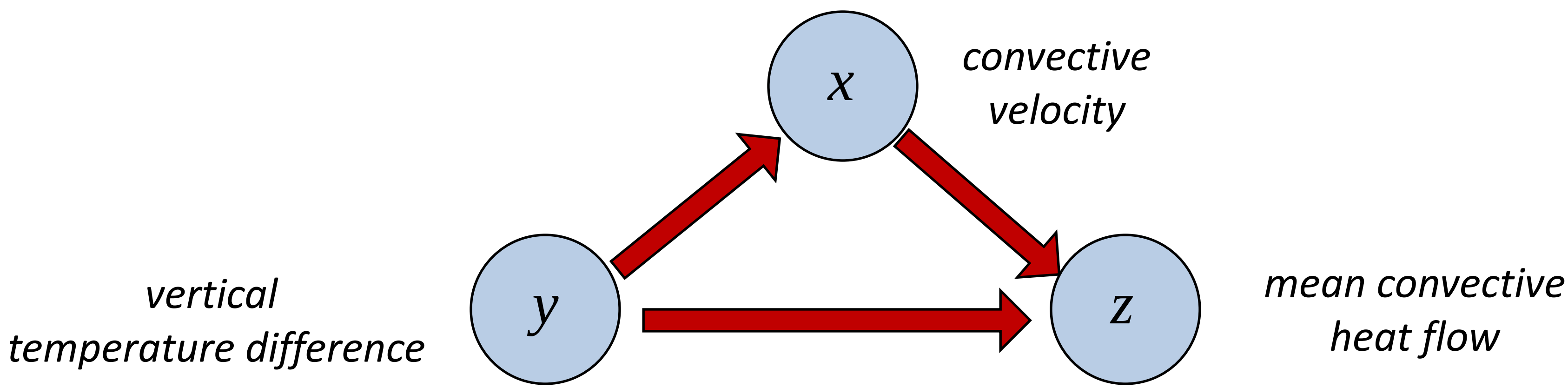
Transfer Entropy Results

Gencaga, Knuth, Rossow
Entropy ***2015***, *17(1), 438-470*

Lorenz system models a two-dimensional convection roll uniformly heated from below and uniformly cooled from above.

x: convective velocity
y: vertical temperature difference
z: mean convective heat flow

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = -xz + rx - y$$
$$\dot{z} = xy - bz$$



Thank You!