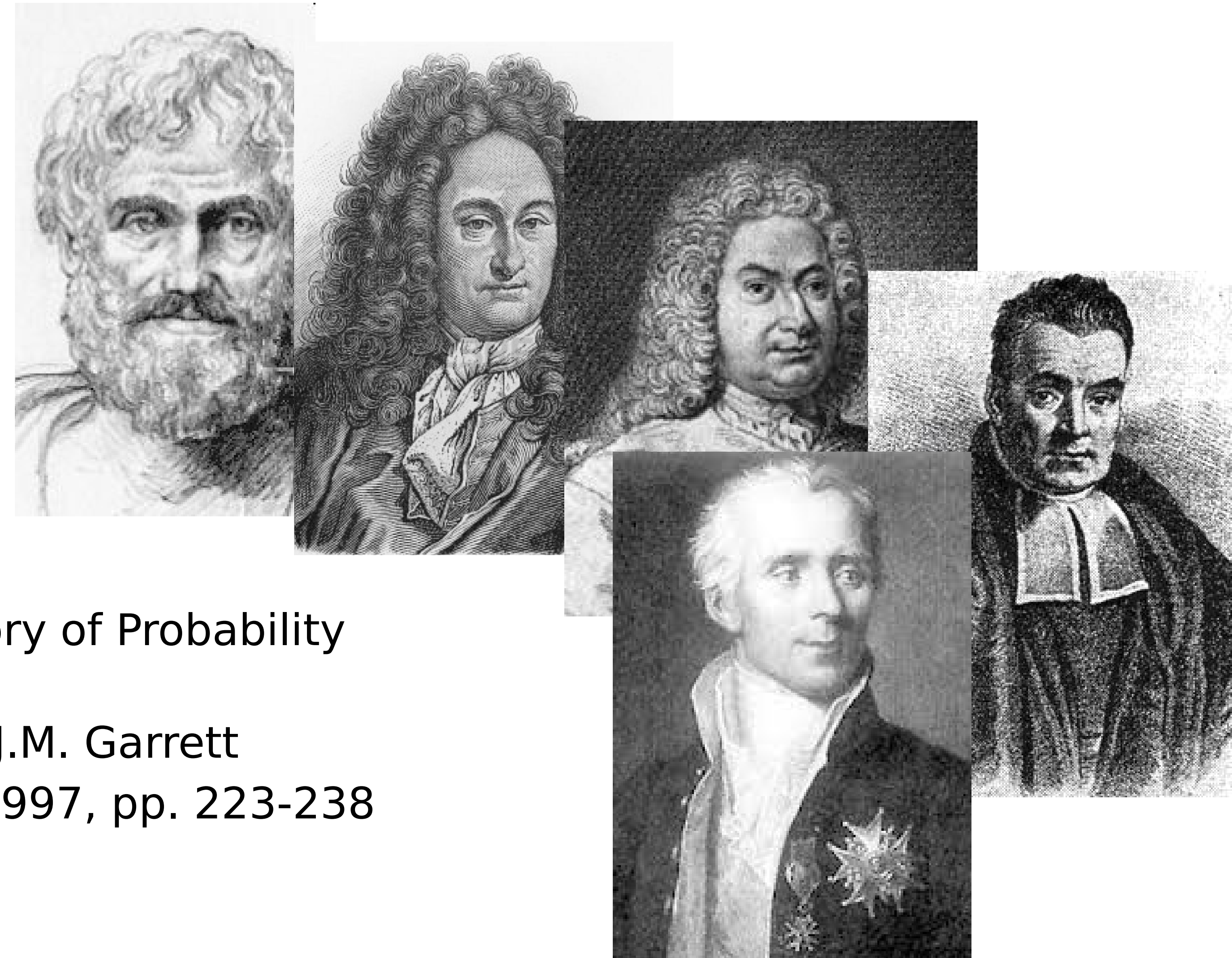
A Modern History of Probability Theory

Kevin H. Knuth
Depts. of Physics and Informatics
University at Albany (SUNY)
Albany NY USA

A Modern History of Probability Theory

Kevin H. Knuth
Depts. of Physics and Informatics
University at Albany (SUNY)
Albany NY USA

A Long History



The History of Probability Theory

Anthony J.M. Garrett

MaxEnt 1997, pp. 223-238

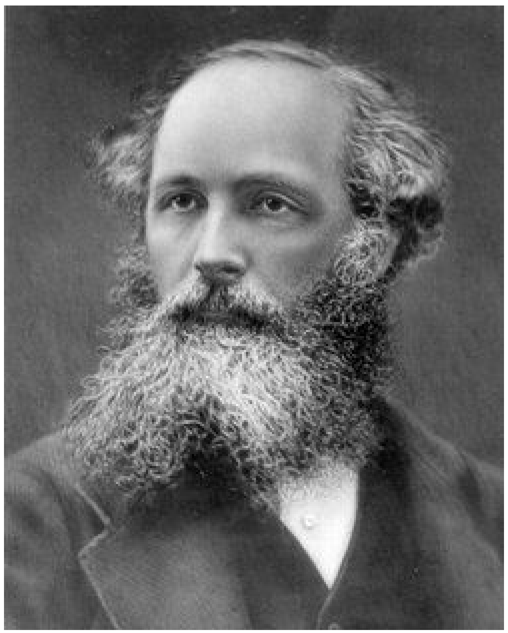
… la théorie des probabilités n'est, au fond
que le bon sens réduit au calcul …

… the theory of probabilities is basically ju
common sense reduced to calculation …

Pierre Simon de Laplace
Théorie Analytique des Probabilités

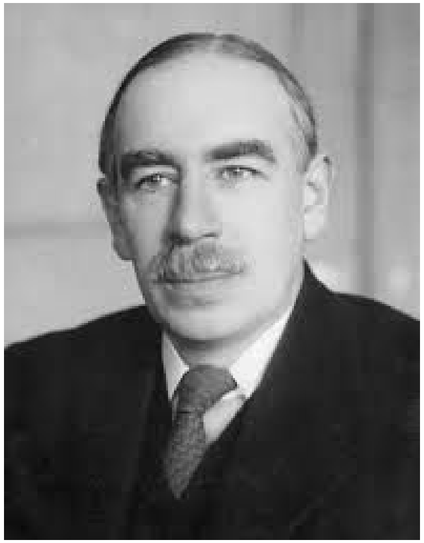
They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

J. CLERK MAXWELL



Taken from Harold Jeffreys "Theory of Probability"

The terms certain and probable describe the various degrees of rational belief about a proposition which different amounts of knowledge authorise us to entertain. All propositions are true or false, but the knowledge we have of them depends on our circumstances; and while it is often convenient to speak of propositions as certain or probable, this expresses strictly a relationship in which they stand to a corpus of knowledge, actual or hypothetical, and not a characteristic of the propositions in themselves. A proposition is capable at the same time of varying degrees of this relationship, depending upon the knowledge to which it is related, so that it is without significance to call a proposition probable unless we specify the knowledge to which we are relating it.



John Maynard Keynes

To this extent, therefore, probability may be called subjective. But in the sense important to logic, probability is not subjective. It is not, that is to say, subject to human caprice. A proposition is not probable because we think it so. When once the facts are given which determine our knowledge, what is probable or improbable in these circumstances has been fixed objectively, and is independent of our opinion. The Theory of Probability is logical, therefore, because it is concerned with the degree of belief which it is rational to entertain in given conditions, and not merely with the actual beliefs of particular individuals, which may or may not be rational.

Meaning of Probability

deriving the laws of probability from more fundamental idea
e has to engage with what ‘probability’ means.

s is a notoriously contentious issue; fortunately, if you disag
h the definition that is proposed, there will be a get-out that
ows other definitions to be preserved.”

- Anthony J.M. Garrett,
“Whence the Laws of Probability”, MaxEnt 1997

Meaning of Probability

The function $p(x|y)$ is often read as 'the probability of x given y'

This is most commonly interpreted as the probability that the proposition x is true given that the proposition y is true.

This concept can be summarized as a **degree of truth**

Concepts of Probability:

- degree of truth

Laplace, Maxwell, Keynes, Jeffreys and Cox all presented a concept of probability based on a **degree of rational belief**.

As Keynes points out, this is not to be thought of as subject to human capriciousness, but rather what an ideally rational agent ought to believe.

Concepts of Probability:

- **degree of truth**
- **degree of rational belief**

Anton Garrett discusses Keynes as conceiving of probability as a **degree of implication**. I don't get that impression reading Keynes. Instead, it seems to me that this is the concept that Garrett had (at the time) adopted.

Garrett uses the word *implicability*.

Concepts of Probability:

- **degree of truth**
- **degree of rational belief**
- **degree of implication**

Concepts of Probability:

- ~~degree of truth~~
- ~~degree of rational belief~~
- degree of implication

Jeffrey Scargle once pointed out that if probability quantifies truth or degrees of belief, one cannot assign a non-zero probability to a model that is known to be an approximation.

One cannot claim to be making inferences with any honesty or consistency while entertaining a concept of probability based on a degree of truth or a degree of rational belief.

Concepts of Probability:

- **degree of truth**
-
-

Jeffrey … truth
or deg… y to a
mode…

One c… esty or
consi… based
on a degree of truth or a degree of rational belief.

Can I give you a “Get-Out”
like Anton did?

Concepts of Probability:

- ~~degree of truth~~
- ~~degree of rational belief within a hypothesis space~~
- degree of implication

Concepts of Probability:

- ~~degree of truth~~
- ~~degree of rational belief~~
- degree of implication

Jeffrey Scargle once pointed out that if probability quantifies truth or degrees of belief, one cannot assign a non-zero probability to a model that is known to be an approximation.

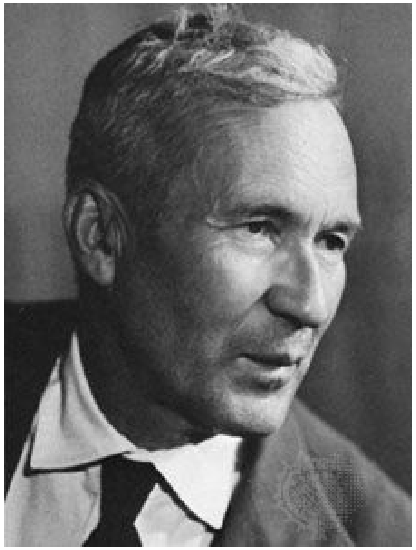
One cannot claim to be making inferences with any honesty or consistency while entertaining a concept of probability based on a degree of truth or a degree of rational belief.

hree Foundations of Probability Theory

Bruno de Finetti - 1931

Foundation Based on Consistent Betting

Unfortunately, the most commonly presented foundation of probability theory in modern quantum foundations



Andrey Kolmogorov - 1933

Foundation Based on Measures on Sets of Events

Perhaps the most widely accepted foundation by modern Bayesians

Richard Threlkeld Cox - 194

Foundation Based on Generalizing Boolean Implication to Degrees

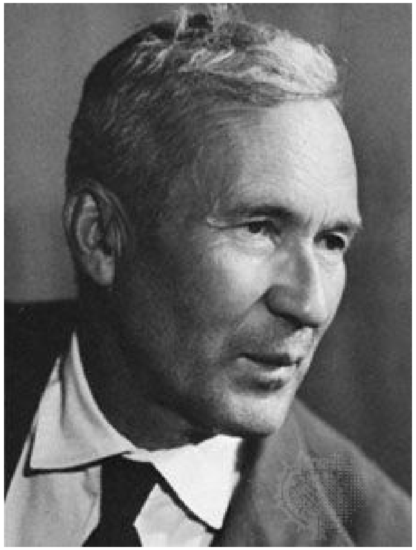
The foundation which has inspired the most investigation and development hree Foundations of Probability Theory

Bruno de Finetti - 1931

Foundation Based on Consistent Betting

Unfortunately, the most commonly presented foundation of probability theory in modern quantum foundations

hree Foundations of Probability Theory



Andrey Kolmogorov - 1933

Foundation Based on Measures on Sets of Events

Perhaps the most widely accepted foundation by modern Bayesians

Axiom I
Probability is quantified by a non-negative real number.

Axiom II
Probability has a maximum value $\Pr(e) \leq 1$ such that the probability that an event in the set E will occur is unity.

Axiom III
Probability is σ-additive, such that the probability of any countable union of disjoint events $e_1, e_2, \cdots \in E$ is given by $\Pr(e_1 \cup e_2 \cup \cdots) = \sum_i^{\infty} \Pr(e_i)$.

Axio Proba r.
Axio Proba the p ur
Axio Proba ny coun giver

It is perhaps the both the conventional nature of his approach and the simplicity of the axioms that has led t such wide acceptance of his foundation.

hree Foundations of Probability Theory

Richard Threlkeld Cox - 1946

Foundation Based on Generalizing Boolean Implication to Degrees

The foundation which has inspired the most investigation and development

Axi
Prol
pro

Axiom 0

Probability quantifies the reasonable credibility of a proposition when another proposition is known to be true

Axi
The

Axiom I

The likelihood $c \cdot b \mid a$ is a function of $b|a$ and $c| \, b \cdot a$

$$c \cdot b \mid a = \mathrm{F}(b|a, c \mid b \cdot a)$$

Axi
The
pro

Axiom II

There is a relation between the likelihood of a proposition and its contradictory

$$\sim b|a = S(b \mid a)$$

In Physics we have a saying,

“The greatness of a scientist is measured by how long he/she retards progress in the field.”

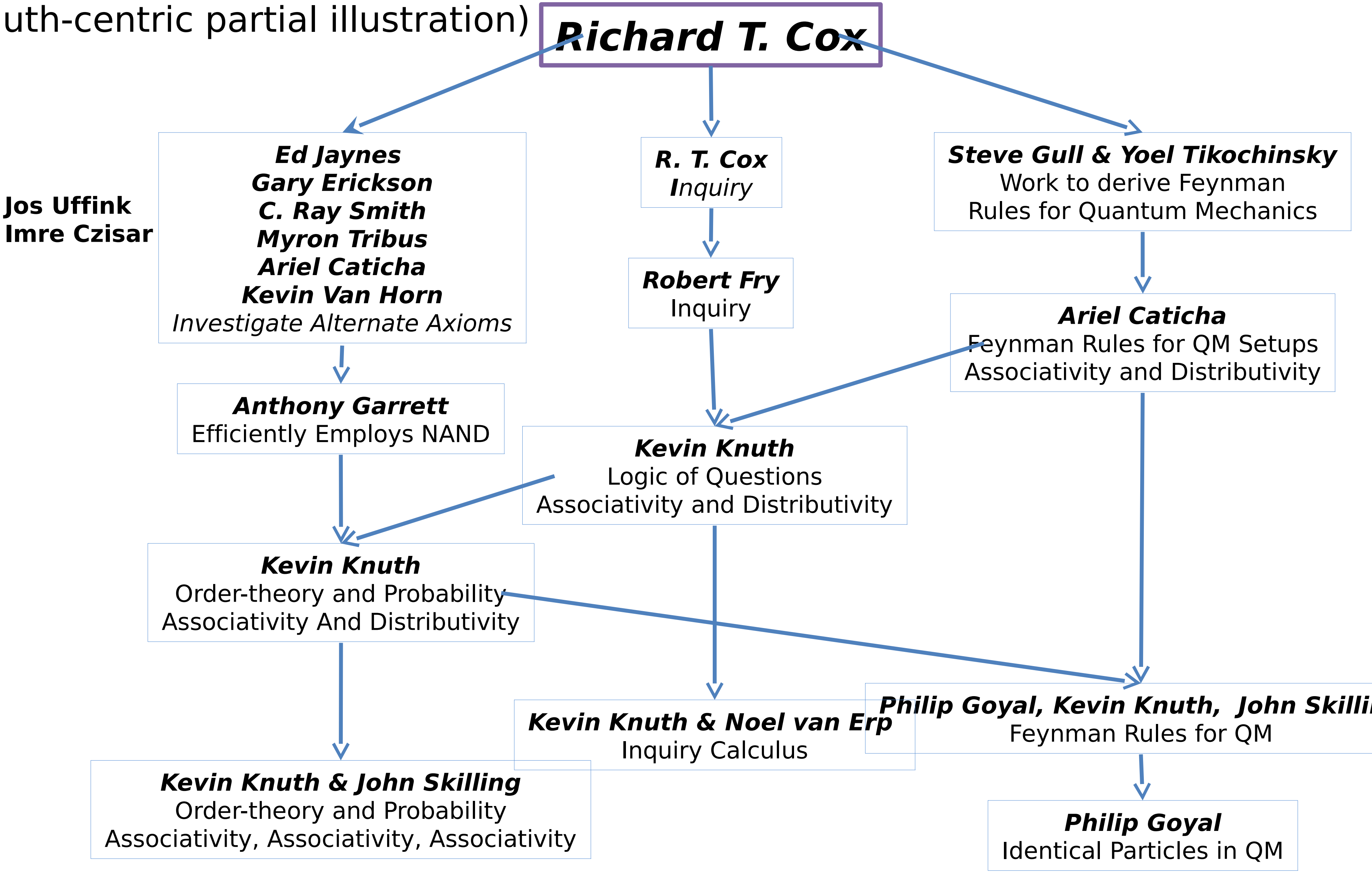
Kolmogorov left few loose ends and no noticeable conceptual glitches to give his disciples sufficient reason or concern to keep investigating.

Cox, on the other hand, proposed a radical approach that raised concerns about how belief could be quantified as well as whether one could improve upon his axioms despite justification by common-sense.

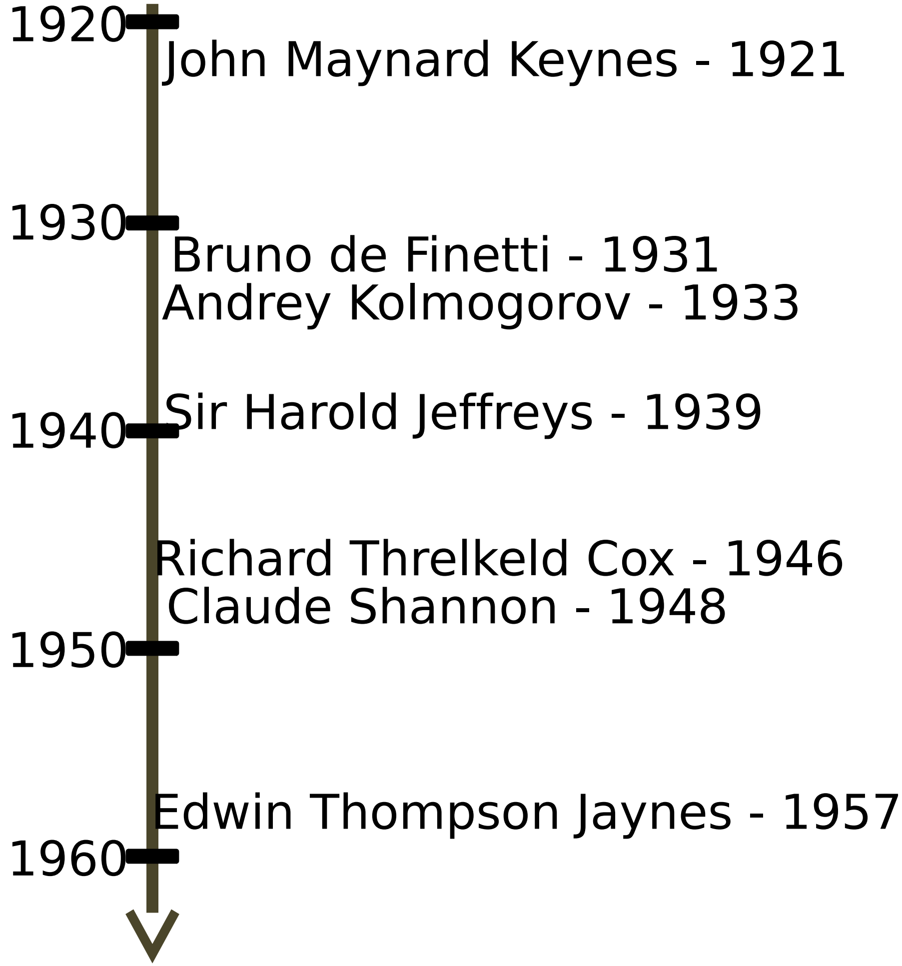
His work was just the right balance between

- Pushing it far enough to be interesting
- Getting it right enough to be compelling
- Leaving it rough enough for there to be remaining work to be done

nd Work Was Done!
Knuth-centric partial illustration)



Probability Theory Timeline



Probability Theory Timeline

1920

John Maynard Keynes - 1921

1930

Bruno de Finetti - 1931

Andrey Kolmogorov - 1933

Sir Harold Jeffreys - 1939

1940

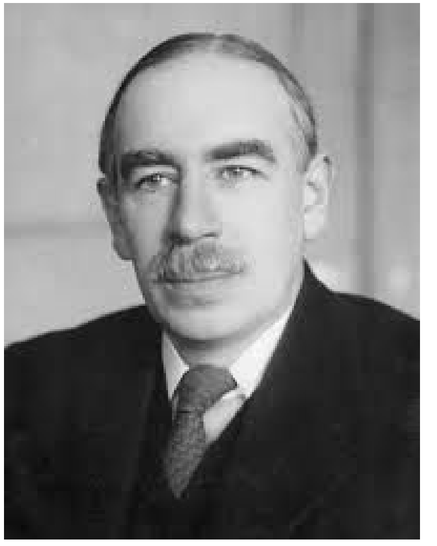
Richard Threlkeld Cox - 1946

Claude Shannon - 1948

1950

Edwin Thompson Jaynes - 1957

1960



Probability Theory Timeline

1920

John Maynard Keynes - 1921

1930

runo de Finetti - 1931

Andrey Kolmogorov - 1933

Sir Harold Jeffreys - 1939

1940

Richard Threlkeld Cox - 1946

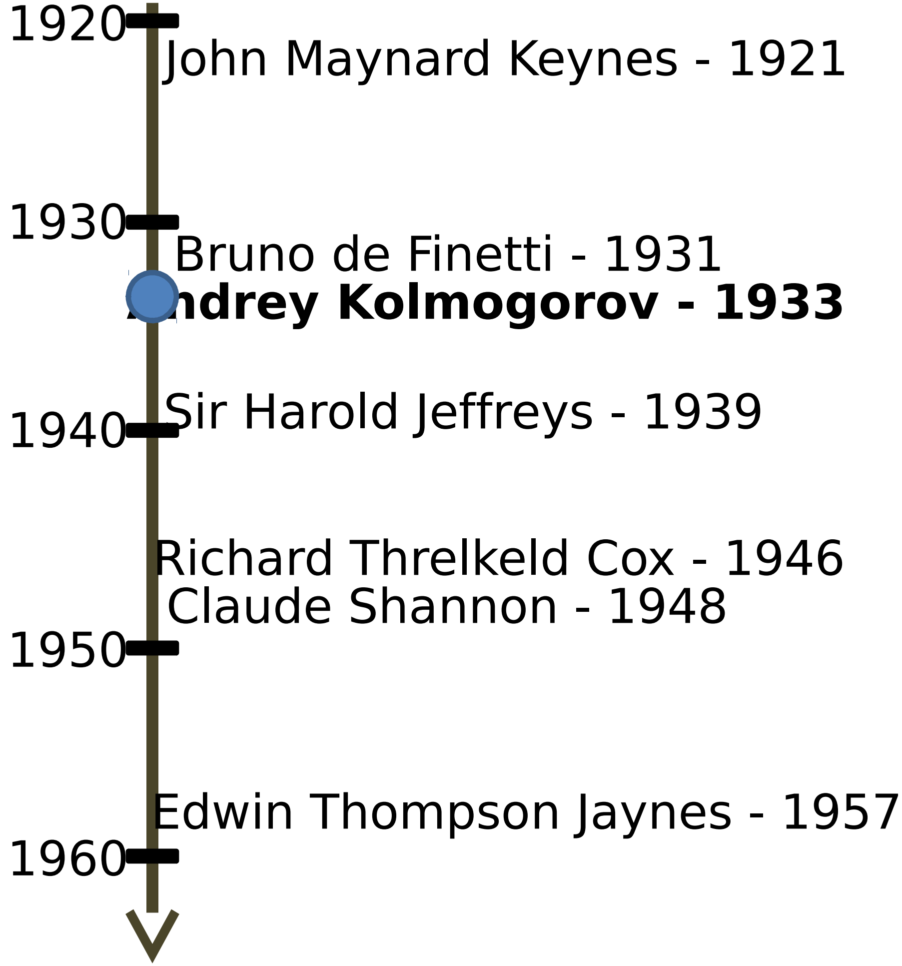
Claude Shannon - 1948

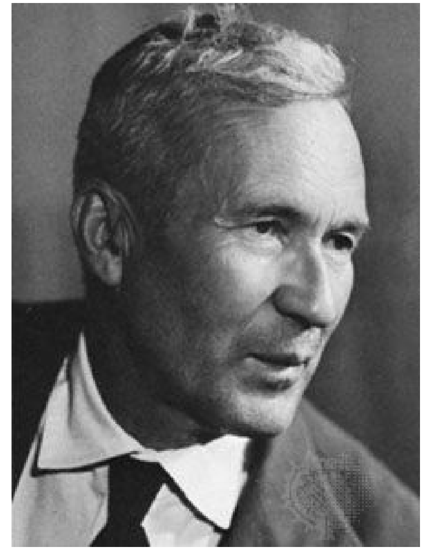
1950

Edwin Thompson Jaynes - 1957

1960

Probability Theory Timeline





Probability Theory Timeline

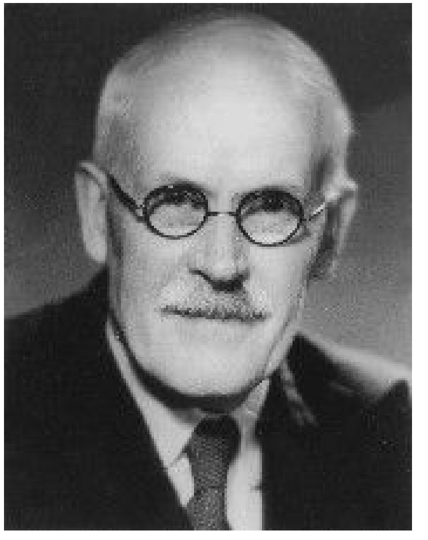
1920
John Maynard Keynes - 1921

1930
Bruno de Finetti - 1931
Andrey Kolmogorov - 1933

r Harold Jeffreys - 1939
1940

Richard Threlkeld Cox - 1946
Claude Shannon - 1948
1950

Edwin Thompson Jaynes - 1957
1960



Probability Theory Timeline

1920
John Maynard Keynes - 1921

1930
Bruno de Finetti - 1931
Andrey Kolmogorov - 1933

Sir Harold Jeffreys - 1939
1940

Richard Threlkeld Cox - 1946
Claude Shannon - 1948
1950

Edwin Thompson Jaynes - 1957
1960

Probability Theory Timeline

1920

John Maynard Keynes - 1921

1930

Bruno de Finetti - 1931

Andrey Kolmogorov - 1933

Sir Harold Jeffreys - 1939

1940

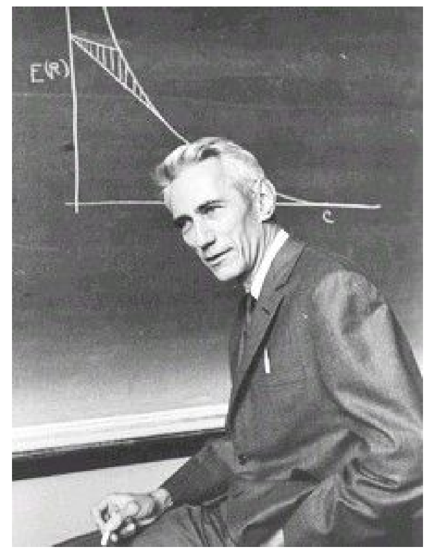
Richard Threlkeld Cox - 1946

laude Shannon - 1948

1950

Edwin Thompson Jaynes - 1957

1960



Probability Theory Timeline

1920

John Maynard Keynes - 1921

1930

Bruno de Finetti - 1931

Andrey Kolmogorov - 1933

Sir Harold Jeffreys - 1939

1940

Richard Threlkeld Cox - 1946

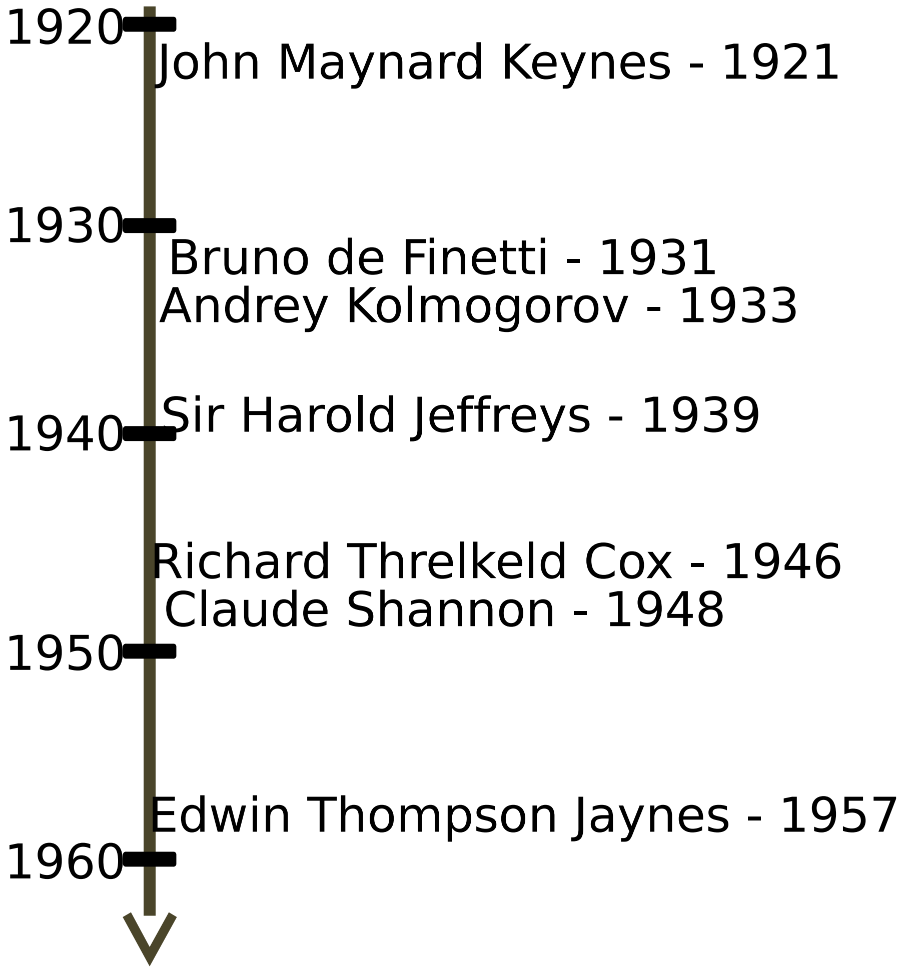
Claude Shannon - 1948

1950

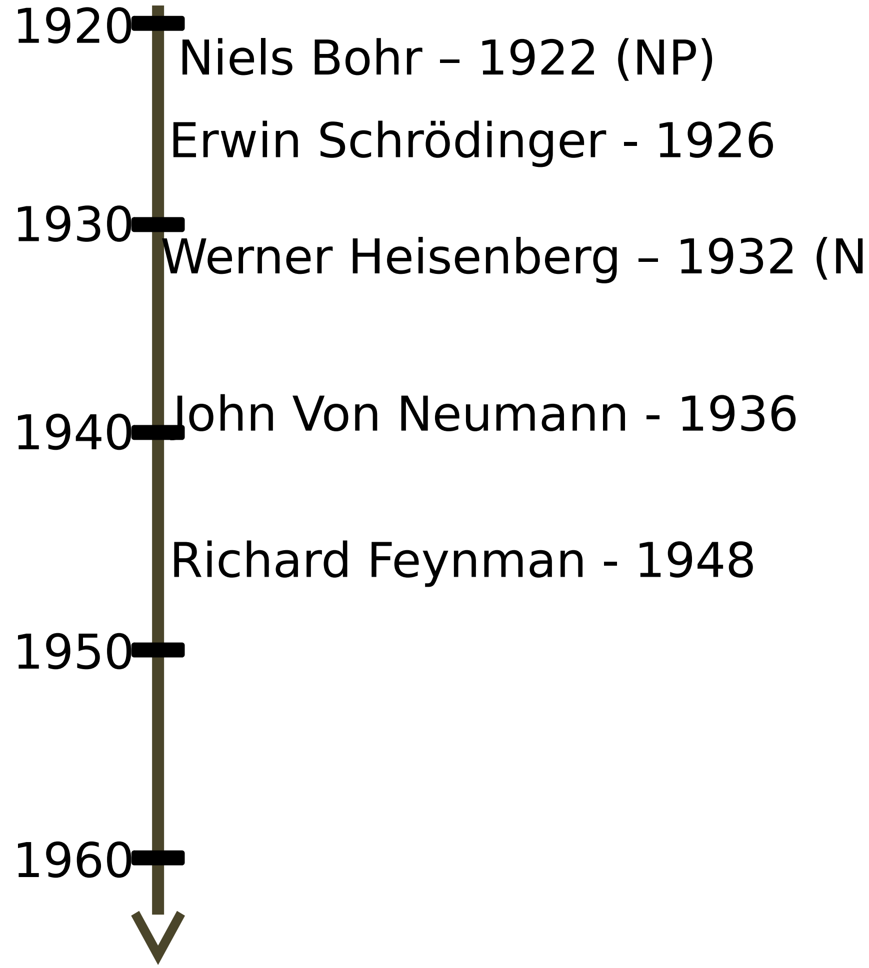
Edwin Thompson Jaynes - 1957

1960

Probability Theory Timeline



Quantum Mechanics Timeline



A Curious Observation

The **Sum Rule for Probability**

$$P(A \text{ or } B \mid C) = P(A \mid C) + P(B \mid C) - P(A \text{ and } B \mid C)$$

Is very much like
the definition of **Mutual Information**

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

However, one cannot be derived from the other.

In fact, the Sum Rule appears to be ubiquitous

$$P(A \text{ or } B \mid C) = P(A \mid C) + P(B \mid C) - P(A \text{ and } B \mid C)$$

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

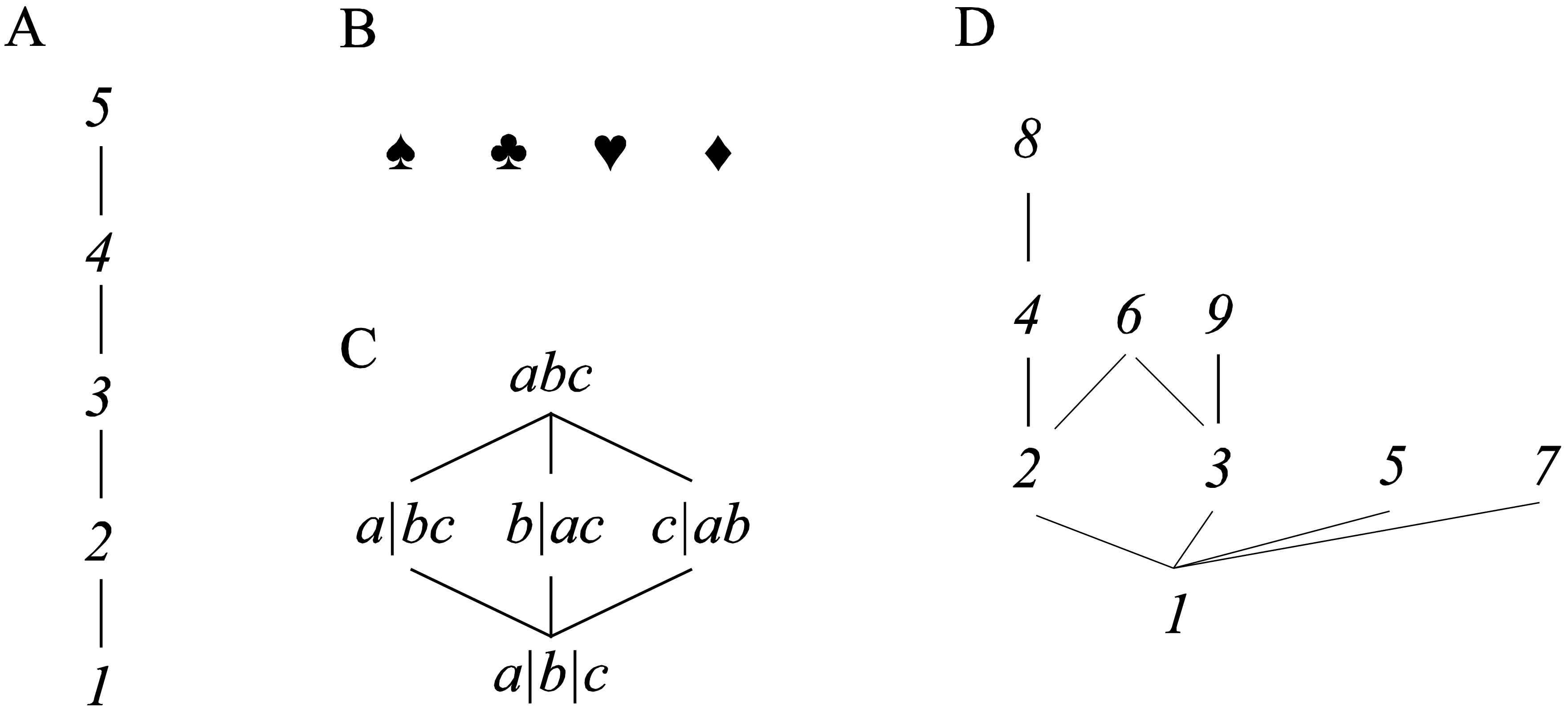
$$m(A \cup B) = m(A) + m(B) - m(A \cap B)$$

$$\chi = F - E + V$$

In Combinatorics the Sum Rule is better known as the **inclusion-exclusion relation**

A MODERN PERSPECTIVE

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

$$a \leq b \quad \Longleftrightarrow \quad \begin{array}{l} a \vee b = b \\ a \wedge b = a \end{array}$$

Lattices

Structural Viewpoint

Operational Viewpoint

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

Sets, Is a subset of

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

Positive Integers, Divides

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

Assertions, Implies

$$a \rightarrow b \iff \begin{matrix} a \vee b = b \\ a \wedge b = a \end{matrix}$$

Integers, Is less than or equal to

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

What can be said about a system?

states

apple

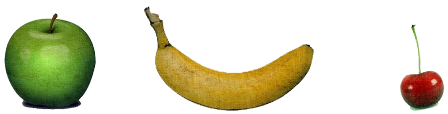
banana

cherry

states of the contents of
my grocery basket

/hat can be said about a system?

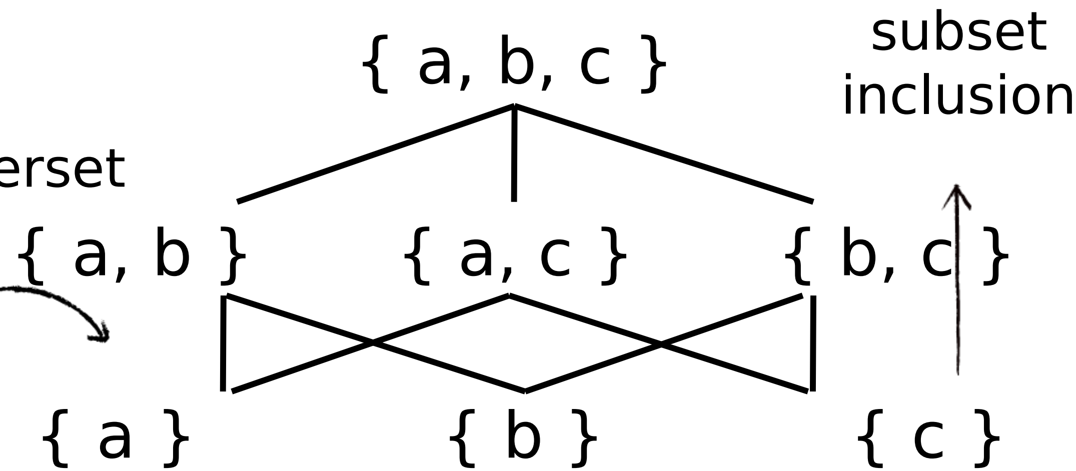
rudely describe knowledge by listing a set of potential states



a b c

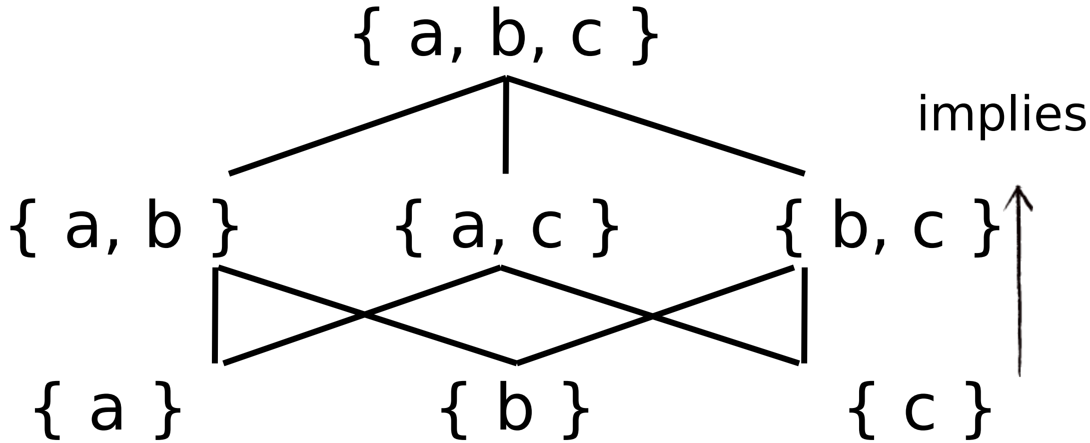
states of the contents of my grocery basket

powerset



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?

{ a, b, c }

{ a, b } { a, c } { b, c }

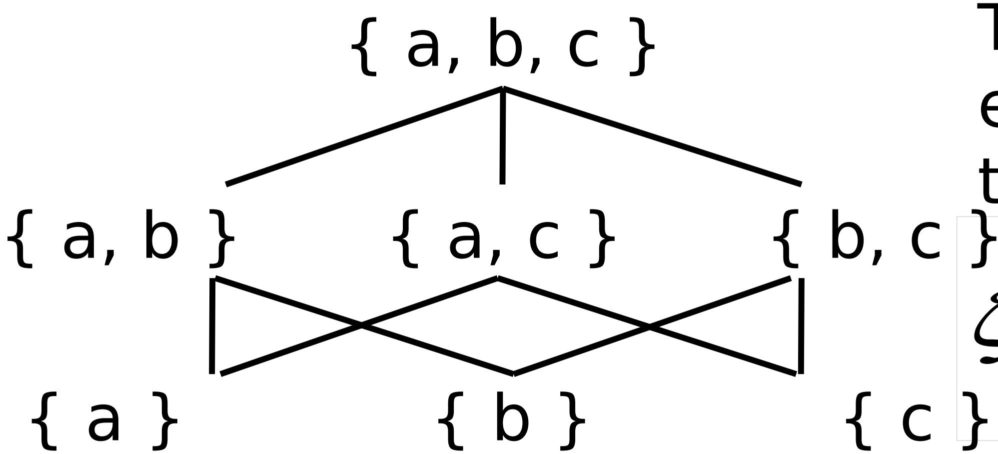
{ a } { b } { c }

statements
about the contents of
my grocery basket

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

inference works backwards

nclusion and the Zeta Function



The Zeta function encodes inclusion on the lattice.

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

The function z

$$z(x,y) = \begin{cases} 1 & \text{if } x \geq y \\ z & \text{if } x \not\geq y \\ 0 & \text{if } x \wedge y = \perp \end{cases}$$

Continues to encode inclusion, but has generalized the concept to degrees of inclusion.

In the lattice of logical statements ordered by implies, this function describes degrees of implication.

The function z

$$z(x,y) = \begin{cases} 1 & \text{if } x \geq y \\ z & \text{if } x \not\geq y \\ 0 & \text{if } x \wedge y = \perp \end{cases}$$

	$\perp$	*a*	*b*	*c*	*a*∨*b*	*a*∨*c*	*b*∨*c*	T
$\perp$	**1**	0	0	0	0	0	0	0
a	**1**	**1**	0	0	?	?	0	?
b	**1**	0	**1**	0	?	0	?	?
c	**1**	0	0	**1**	0	?	?	?
a∨*b*	**1**	**1**	**1**	0	**1**	?	?	?
a∨*c*	**1**	**1**	0	**1**	?	**1**	?	?
b∨*c*	**1**	0	**1**	**1**	?	?	**1**	?
T	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Are all of the values of the function z arbitrary?

Or are there constraints?

Probability

Changing notation

$$z(x,y) = \begin{cases} 1 & \text{if } x \geq y \\ z & \text{if } x \not\geq y \\ 0 & \text{if } x \wedge y = \perp \end{cases}$$

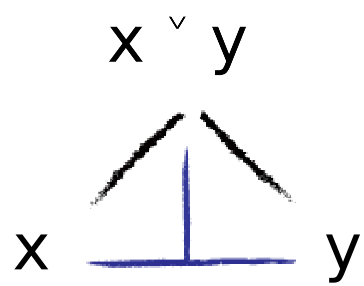
$$P(x\,|\,y) = \begin{cases} 1 & \text{if } y \rightarrow x \\ 0 < p < 1 & \text{if } y \not\rightarrow x \\ 0 & \text{if } x \wedge y = \perp \end{cases}$$

The MEANING of $P(x|y)$ is made explicit via the Zeta function.

These are degrees of implication!

VALUATION $v : x \in L \to \mathbb{R}$

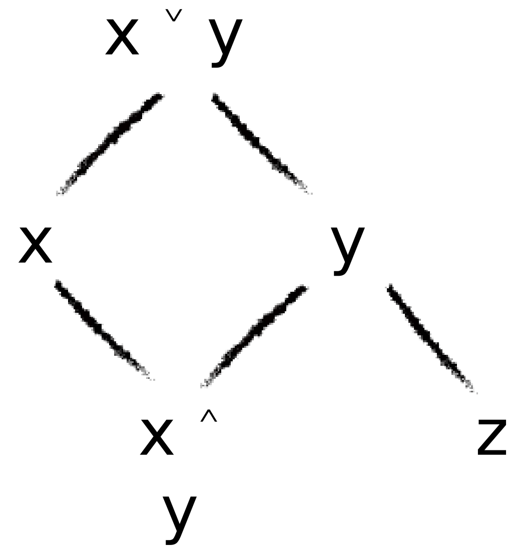
If $y \geq x$ then $v(y) \geq v(x)$



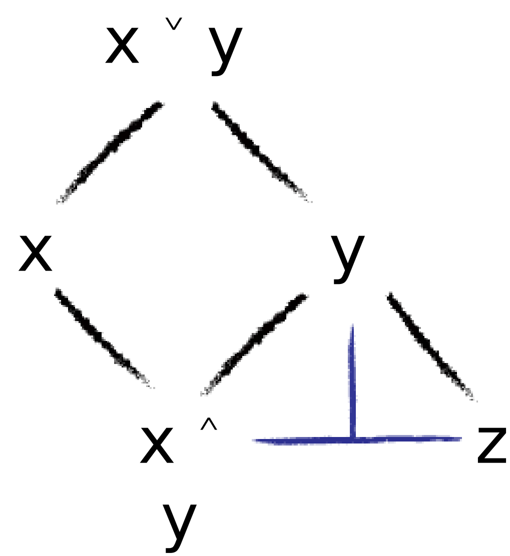
$$v(x \vee y) = v(x) + v(y)$$

Associativity and Order implies Additivity
(up to arbitrary invertible transform)

General Case

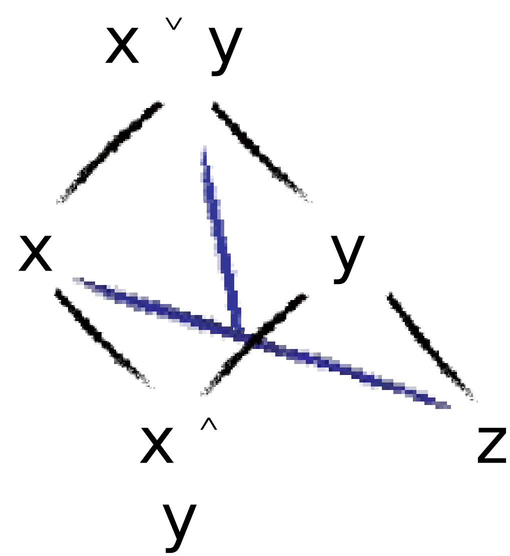


General Case



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

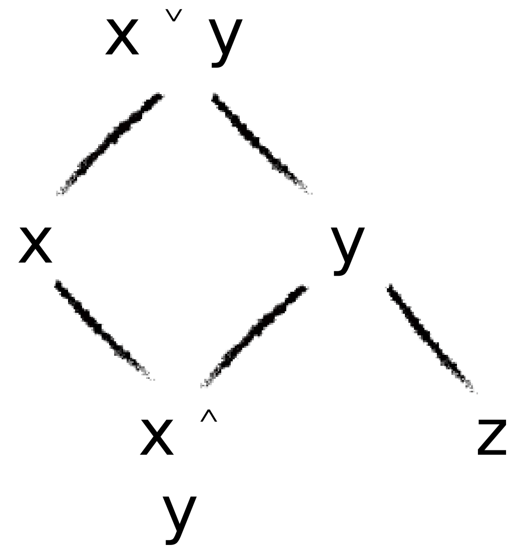
General Case



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

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

General Case



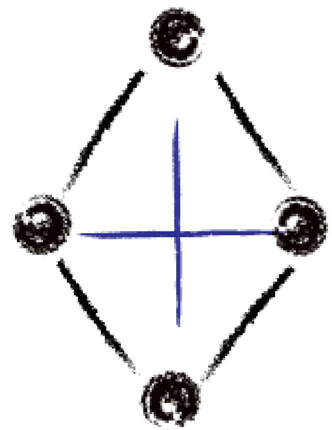
$$v(y) = v(x \wedge y) + v(z)$$

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

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

Sum Rule

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



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

symmetric form (self-dual)

Sum Rule

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

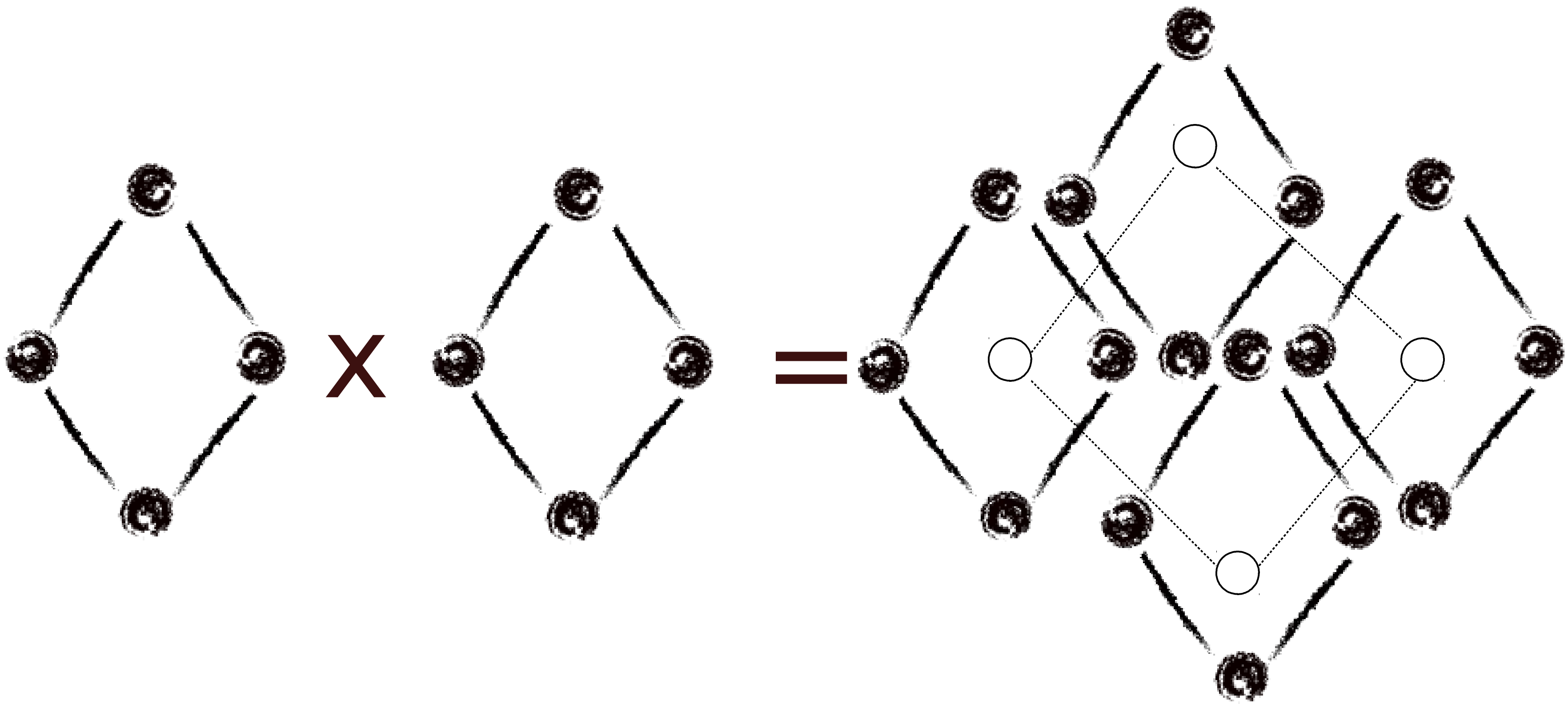
$$MI(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$\max(x, y) = x + y - \min(x, y)$$

$$\chi = V - E + F$$

$$\log(\gcd(x, y)) = \log(x) + \log(y) - \log(\mathrm{lcm}(x, y))$$

Lattice Products



Direct (Cartesian) product of two spaces

Direct Product Rule

The lattice product is also associative

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

After the sum rule, the only freedom left is rescaling

$$v((a, b)) = v(a)\, v(b)$$

which is again summation (after taking the logarithm)

Context and Bi-Valuations

BI-VALUATION $w : x, i \in L \to \mathbb{R}$

Bi-Valuation $w(x \mid i)$

$v_i(x)$

Valuation $v(x)$

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

Context is Explicit

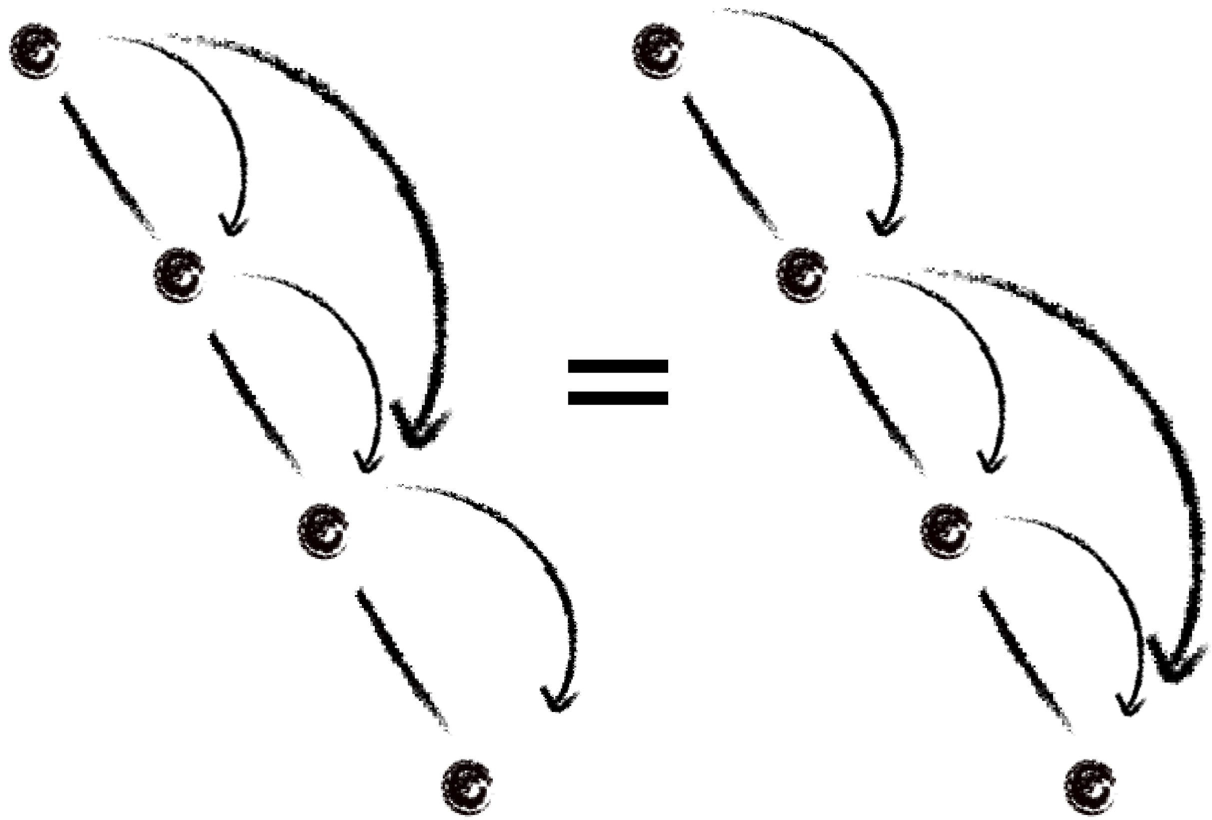
Sum Rule

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

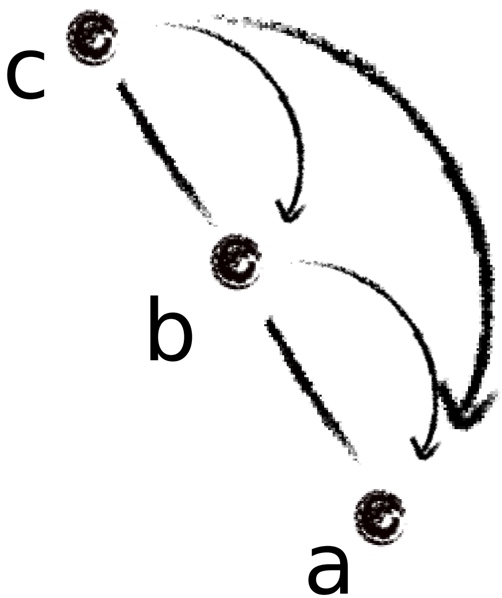
Direct Product Rule

$$w((a, b) \mid (i, j)) = w(a \mid i)\, w(b \mid j)$$

Associativity of Context



Quantifying Lattices



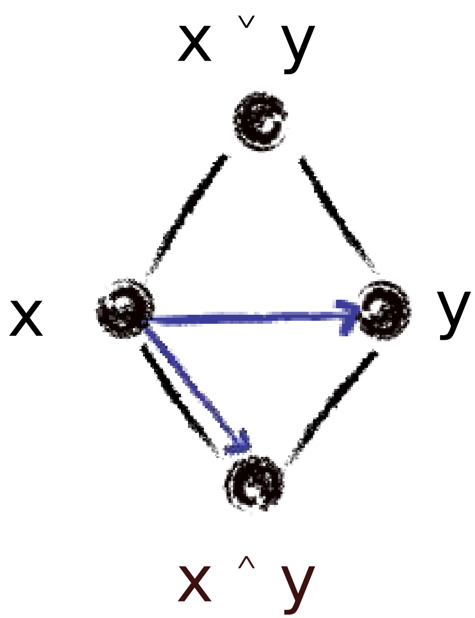
Chain Rule

$$w(a|c) = w(a|b)\, w(b|c)$$

Lemma

$$w(x\,|\,x) + w(y\,|\,x) = w(x \vee y\,|\,x) + w(x \wedge y\,|\,x)$$

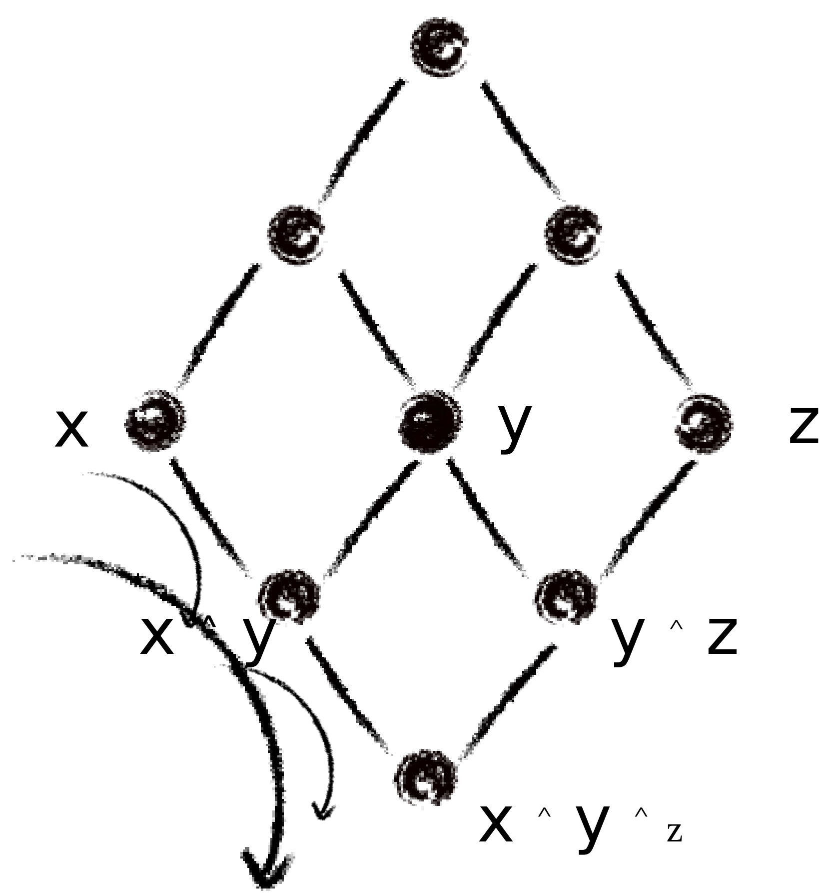
Since $x \leq x$ and $x \leq x \vee y$, $w(x\,|\,x) = 1$ and $w(x \vee y\,|\,x) =$



$$w(y\,|\,x) = w(x \wedge y\,|\,x)$$

Extending the Chain Rule

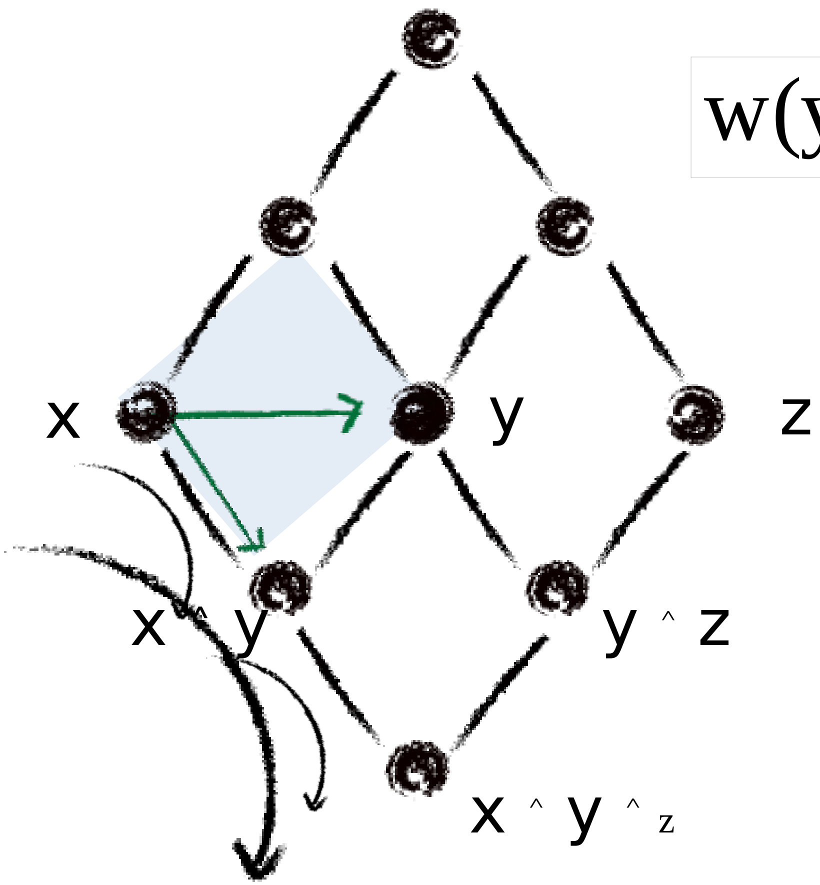
$$w(x \wedge y \wedge z \mid x) = w(x \wedge y \mid x)\, w(x \wedge y \wedge z \mid x \wedge y)$$



Extending the Chain Rule

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

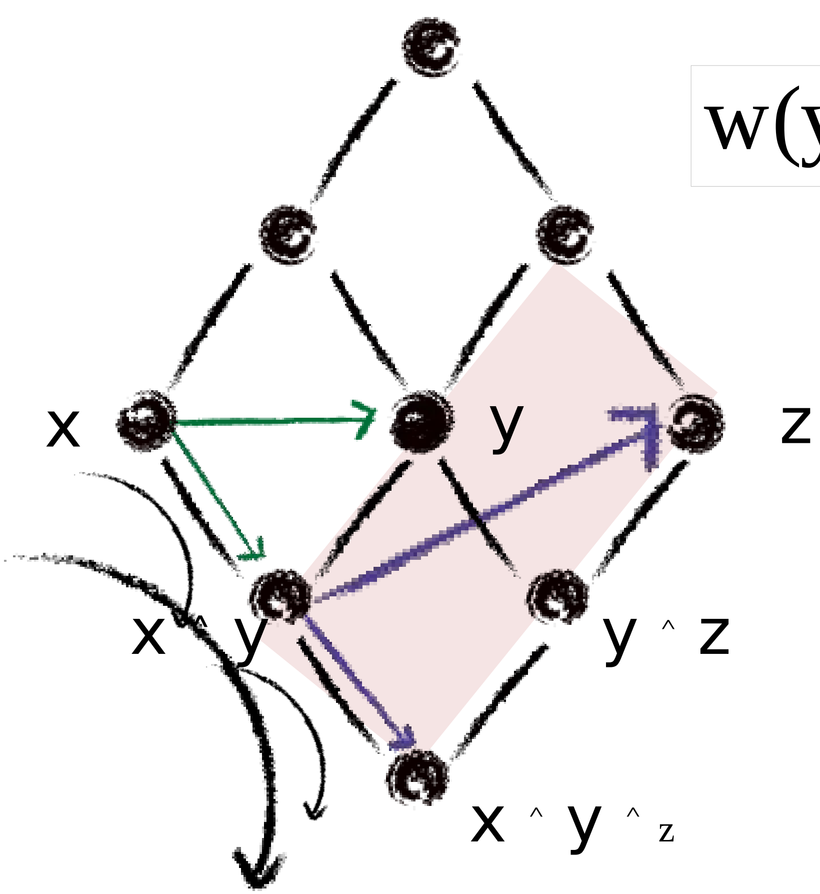
$$w(y \wedge z \mid x) = w(y \mid x)\,w(z \mid x \wedge y)$$



Extending the Chain Rule

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

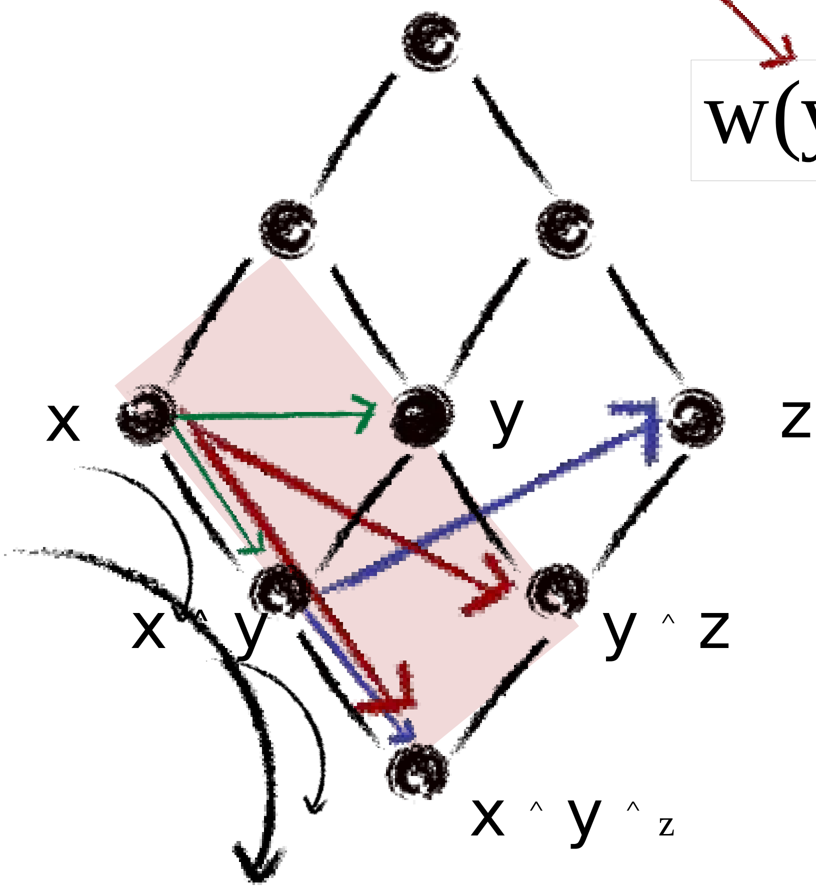
$$w(y \wedge z \mid x) = w(y \mid x)\, w(z \mid x \wedge y)$$



Extending the Chain Rule

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

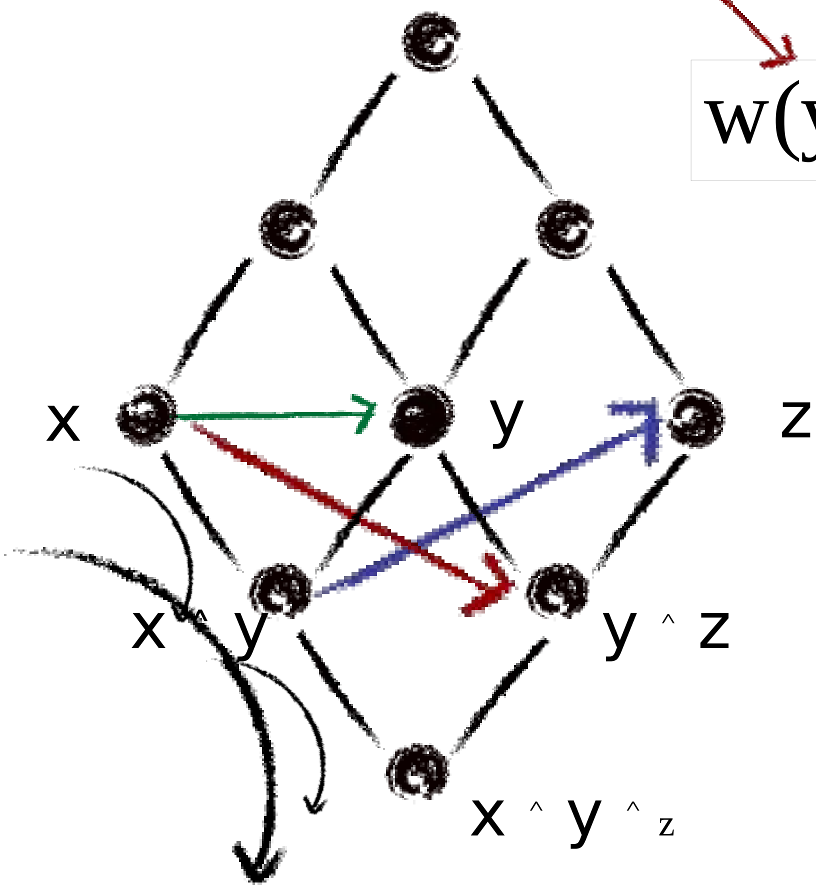
$$w(y \wedge z \mid x) = w(y \mid x)\,w(z \mid x \wedge y)$$



Extending the Chain Rule

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

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



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

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

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

Bayes Theorem involves a change of context.

Bayesian Probability Theory
Constraint Equations

Sum Rule

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

Direct Product Rule

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

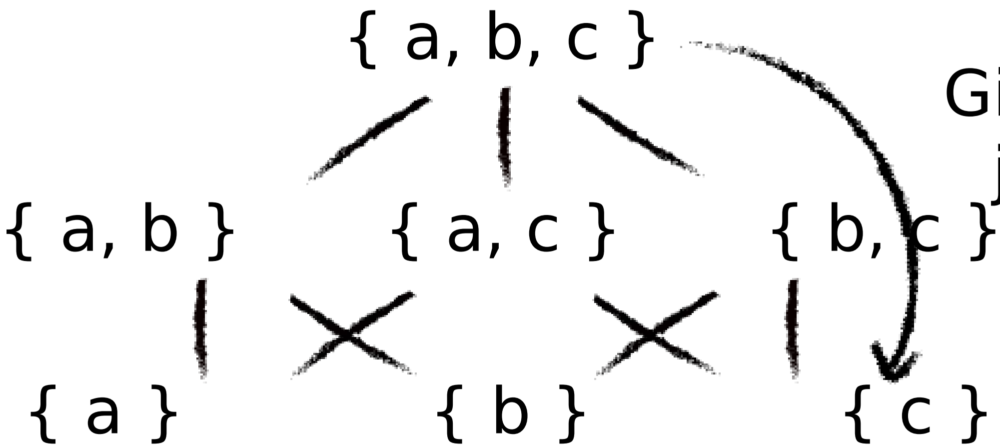
Product Rule

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

Bayes Theorem

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

Inference

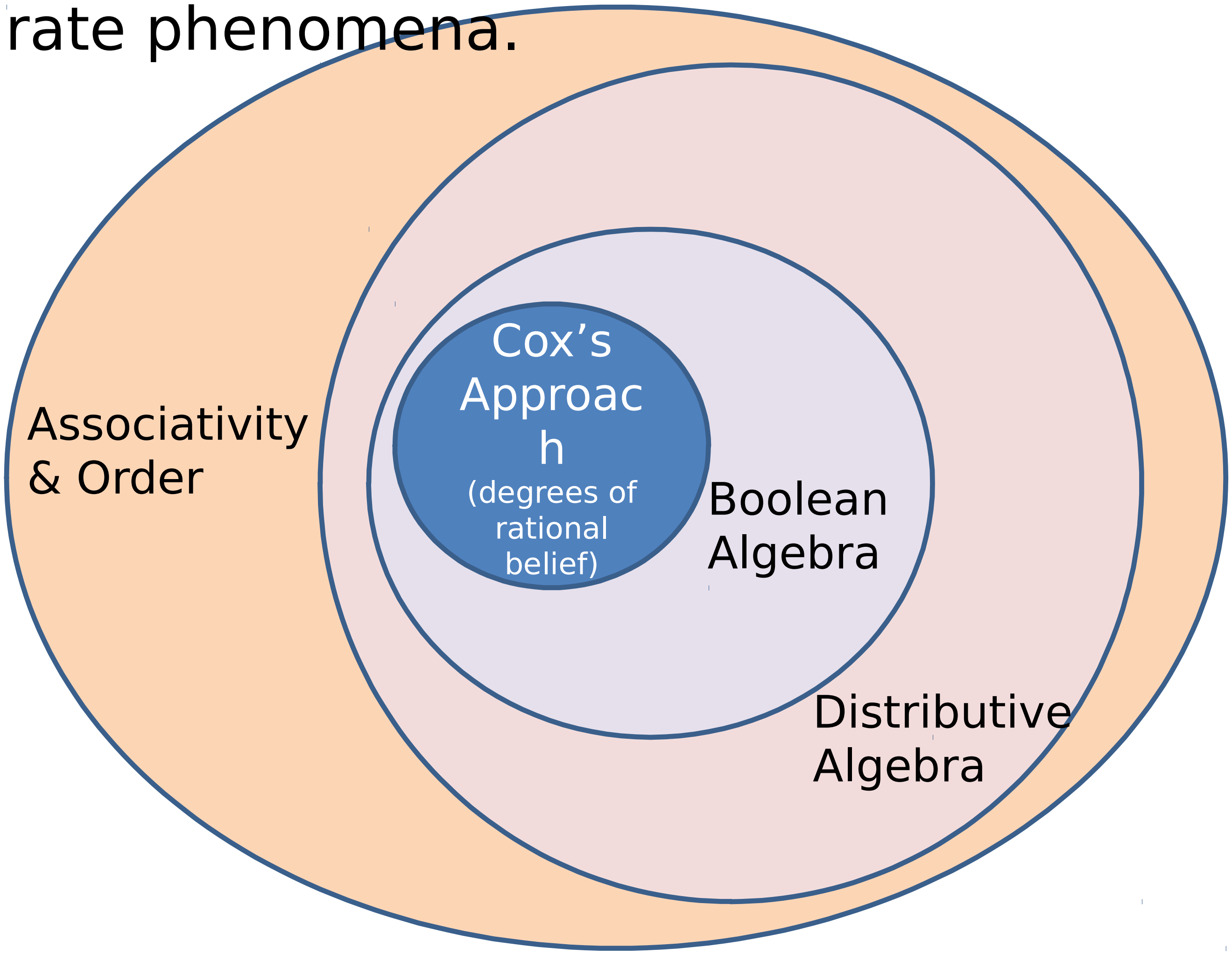


statements

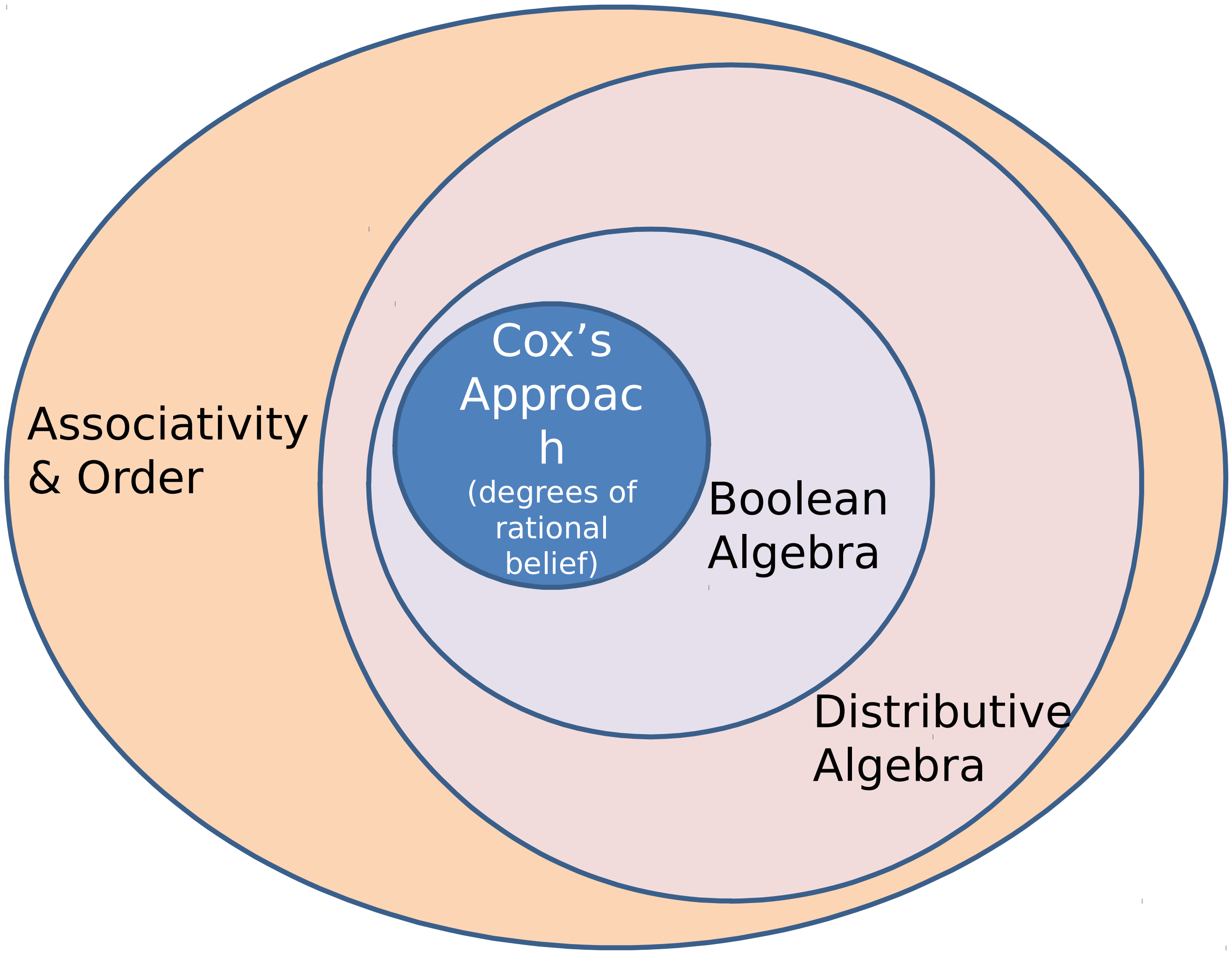
Given a quantification of the join-irreducible elements, one uses the constraint equations to consistently assign any desired bi-valuations (probability)

Foundations are Important.

A solid foundation acts as a broad base on which theories can be constructed to unify seemingly disparate phenomena.

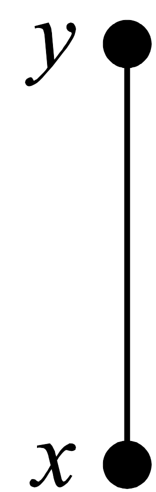


THANK YOU


Associativity
& Order
Cox’s
Approac
h
(degrees of
rational
belief)
Boolean
Algebra
Distributive
Algebra

Quantification of a Lattice

To constrain the form of the function f where $d = f(a, b, c)$ consider the chain given by $x \leq y$.



Since x and y are totally ordered we have that

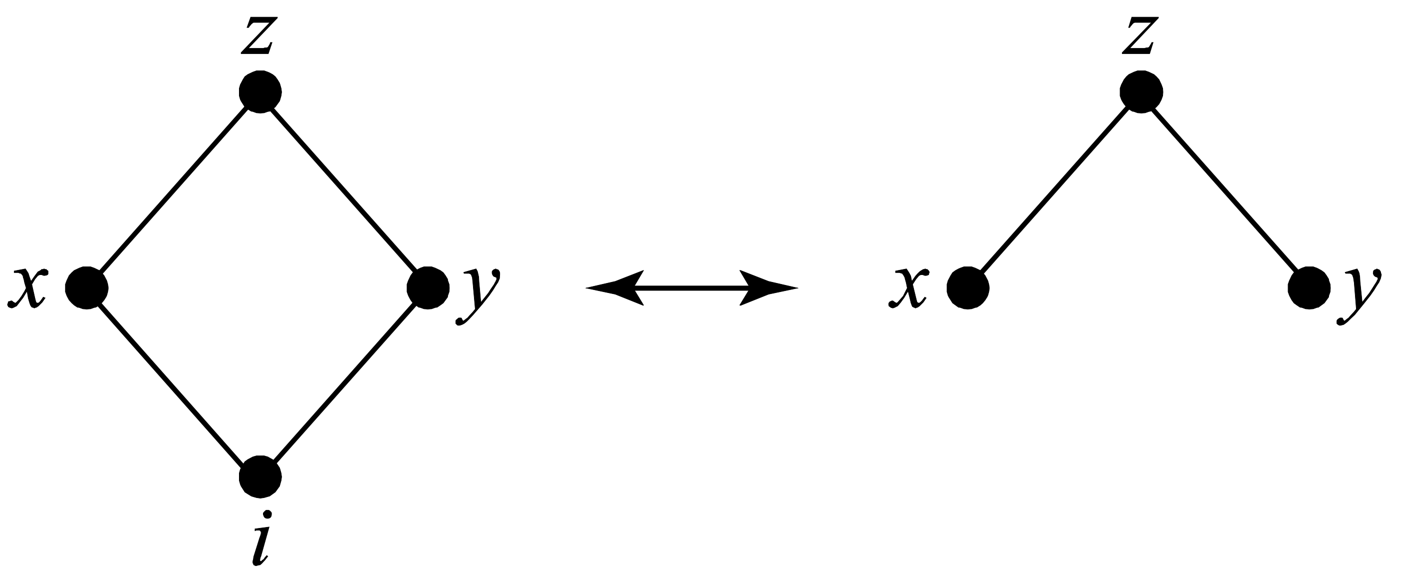
$$y = f(x, y, x)$$

and

$$y = f(y, x, x)$$

by commutativity.

Some lattices are drawn as semi-join lattices where the bottom element is optional

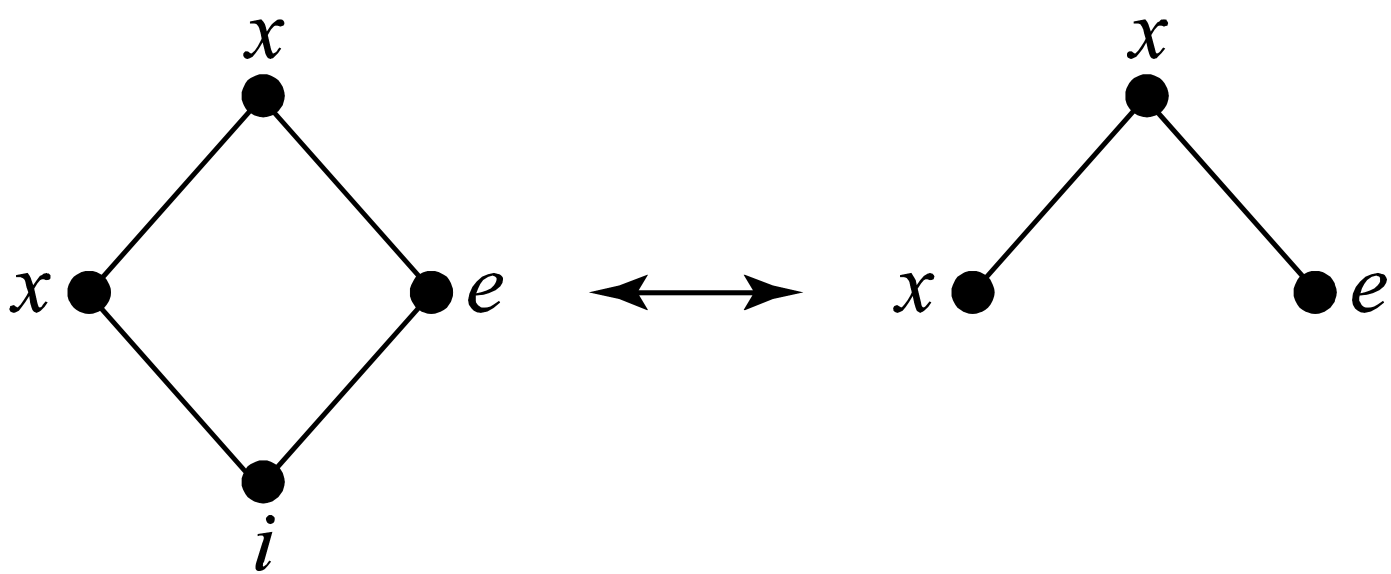


$$z = f(x, y, i) \doteq x \oplus y$$

where $\oplus$ is an real-valued operator to be determined.

Quantification of a Lattice

Consider the identity quantification e, where $x \oplus e = x$



This implies that $f(x, e, x) = i$

Given the chain result: $y = f(x, y, x)$

We have that $f(x, e, x) = e = i$ so that the optional bottom is assigned the $\oplus$-identity.

Quantification of a Lattice

We have that

$$z = f(y, e, x)$$

Also

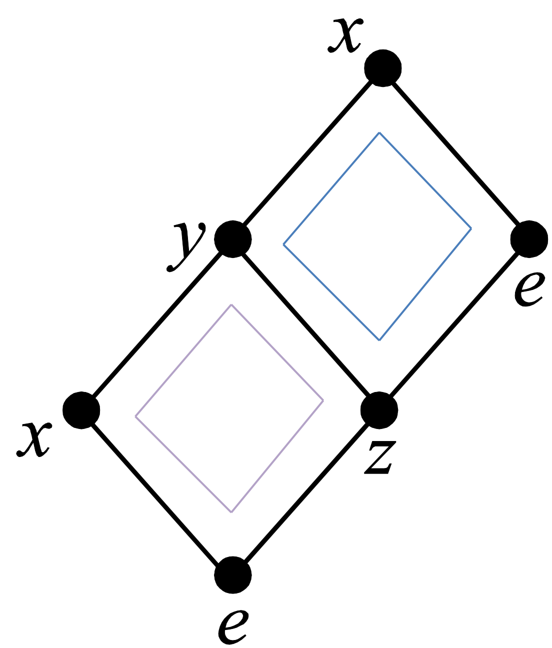
$$y = x \oplus z$$

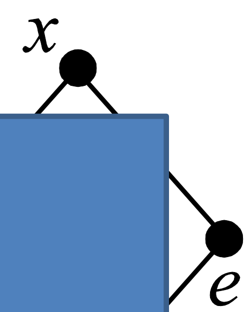
So that

$$z = f(x \oplus z\,, e, x)$$

Rewriting $f(x \oplus z\,, e, x) \equiv g(x \oplus z\,, x)$
we have that

$$z = g(x \oplus z\,, x)$$
$$\equiv (x \oplus z) \ominus x$$





We have that

$\boxed{z = f(y, e, x)}$

Also

$\boxed{y = x \oplus z}$

So that

$z = f(x \oplus z\ , e$

Rewriting $f(x \oplus$

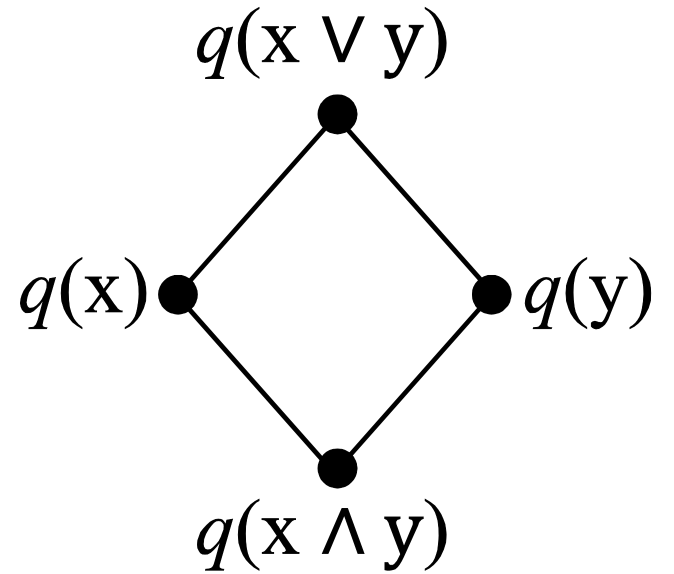
we have that

$$z = g(x \oplus z\ , x)$$
$$\equiv (x \oplus z) \ominus x$$

Sum Rule

Given that $\oplus$ is commutative and associative, we have that it is Abelian.

$$q(\mathrm{x} \vee \mathrm{y}) = q(\mathrm{x}) \oplus q(\mathrm{y}) \ominus q(\mathrm{x} \wedge \mathrm{y})$$



One can then show that in the case of valuations, $\oplus$ is an invertible transform of the usual addition
(eg. Craigen & Pales 1989; Knuth & Skilling 2012)

$$q(\mathrm{x} \vee \mathrm{y}) = q(\mathrm{x}) + q(\mathrm{y}) - q(\mathrm{x} \wedge \mathrm{y})$$