Automating Science

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

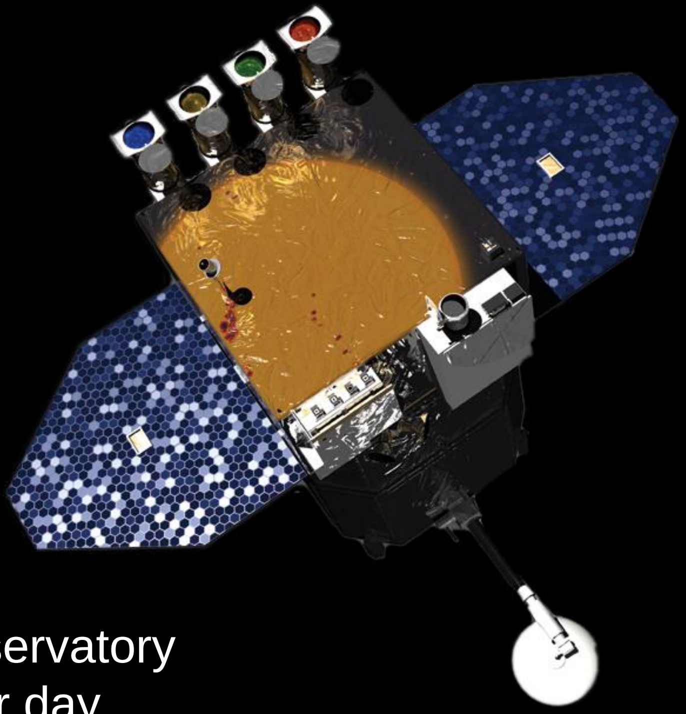
Supported by:
NASA Applied Information Systems Research Program (AISRP)
NASA Applied Information Systems Technology Program (AIST)

Massive Data Collection

NASA Earth Observatories
Jason-1
Landsat 7
QuikSCAT
ERBS
ACRIMSAT
Aqua
Terra
TOPEX/Poseidon
TRMM
SAGE III/METEOR-3M
UARS
SORCE
GRACE
SeaWiFS
EP-TOMS
NMP/EO-1
ICESat
Aura

3 Terabytes of data per day.
Storage approaching 10 Petabytes

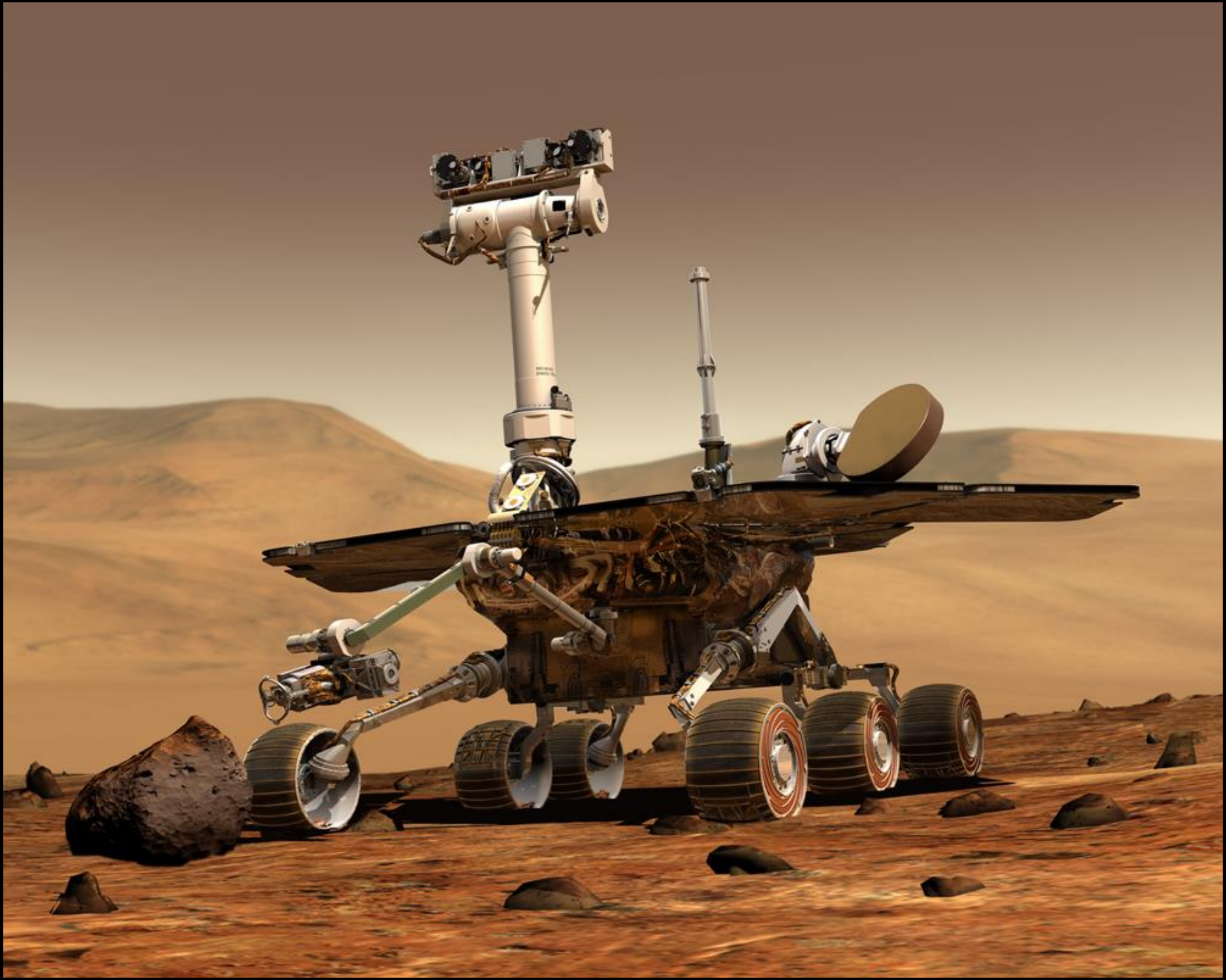
Massive Data Collection



Solar Dynamics Observatory
1.5 Terabytes per day
0.75 Petabytes per year

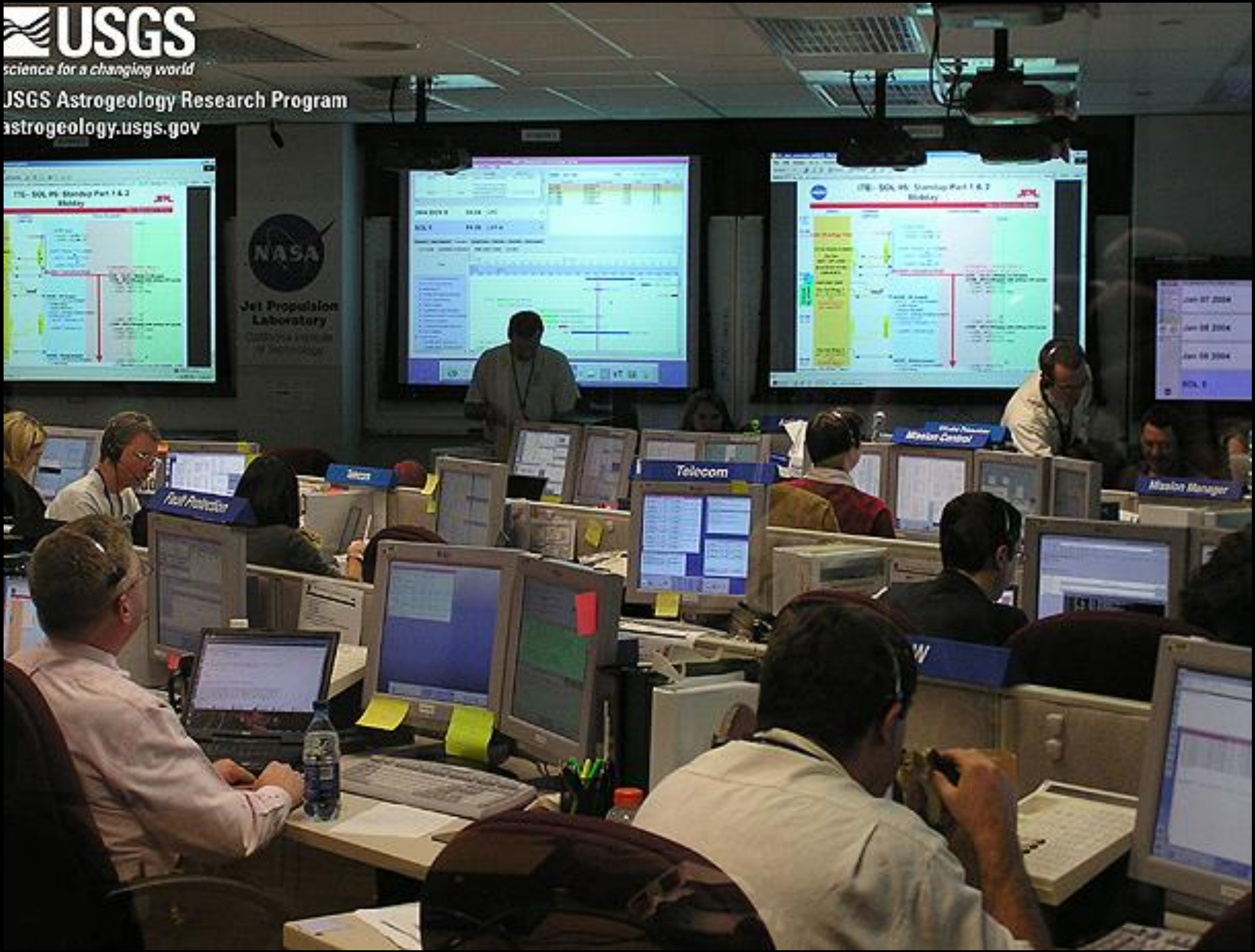
The Data Fire Hose

Focused Exploration



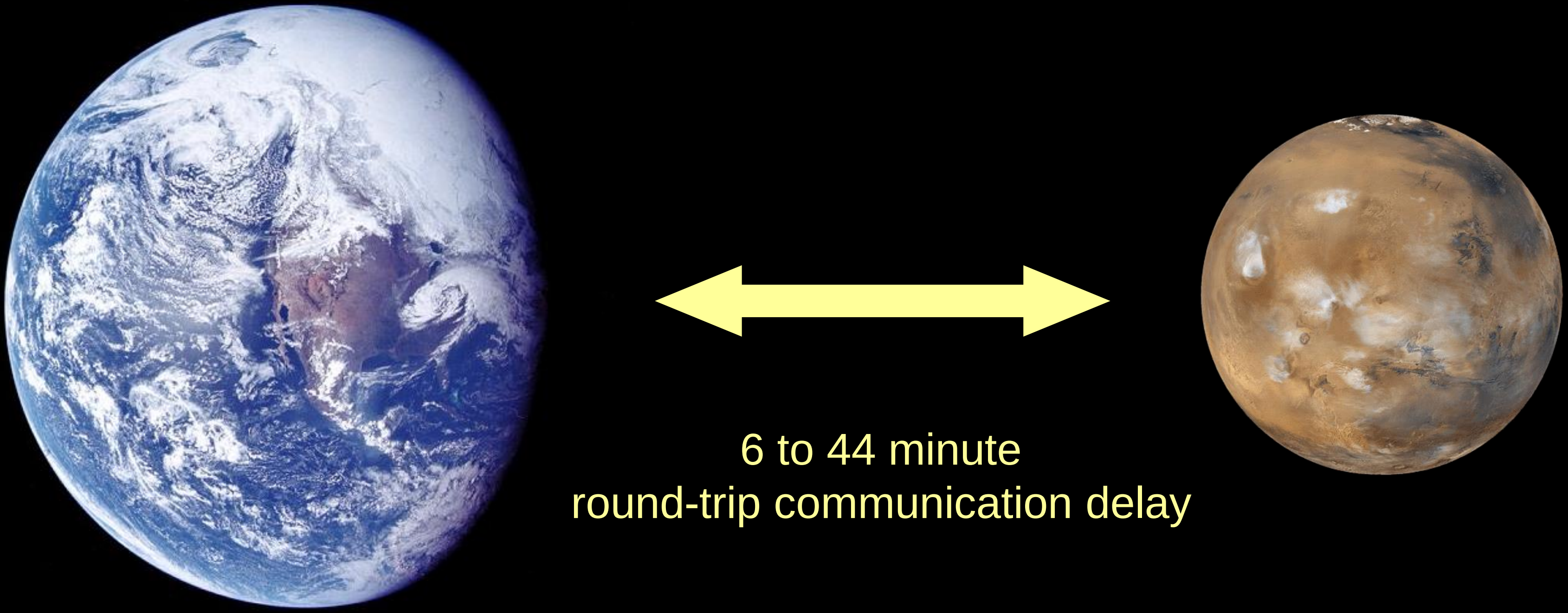
Mars Exploration Rovers: Spirit and Opportunity
128 kilobits per second / 10 Megabytes per day

Mars Exploration Rover Mission Control


USGS
science for a changing world
USGS Astrogeology Research Program
astrogeology.usgs.gov
NASA
Jet Propulsion Laboratory
California Institute of Technology
Telecom
Fault Protection
Mission Control
Mission Manager

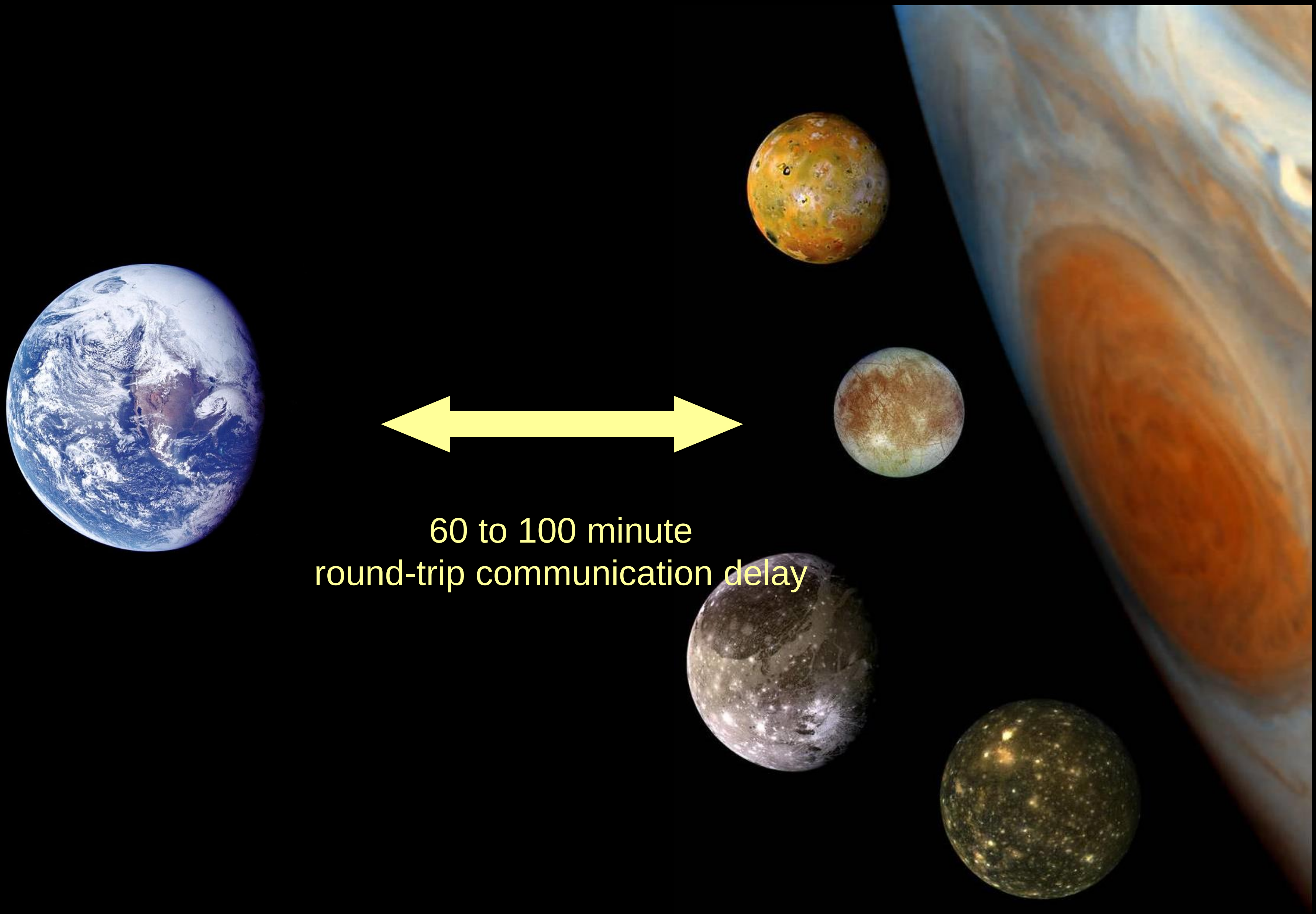
Event: MER Mission Activities
Date: Spirit Sol 4
Source: Kris Becker

Time Constraints and Human Intervention



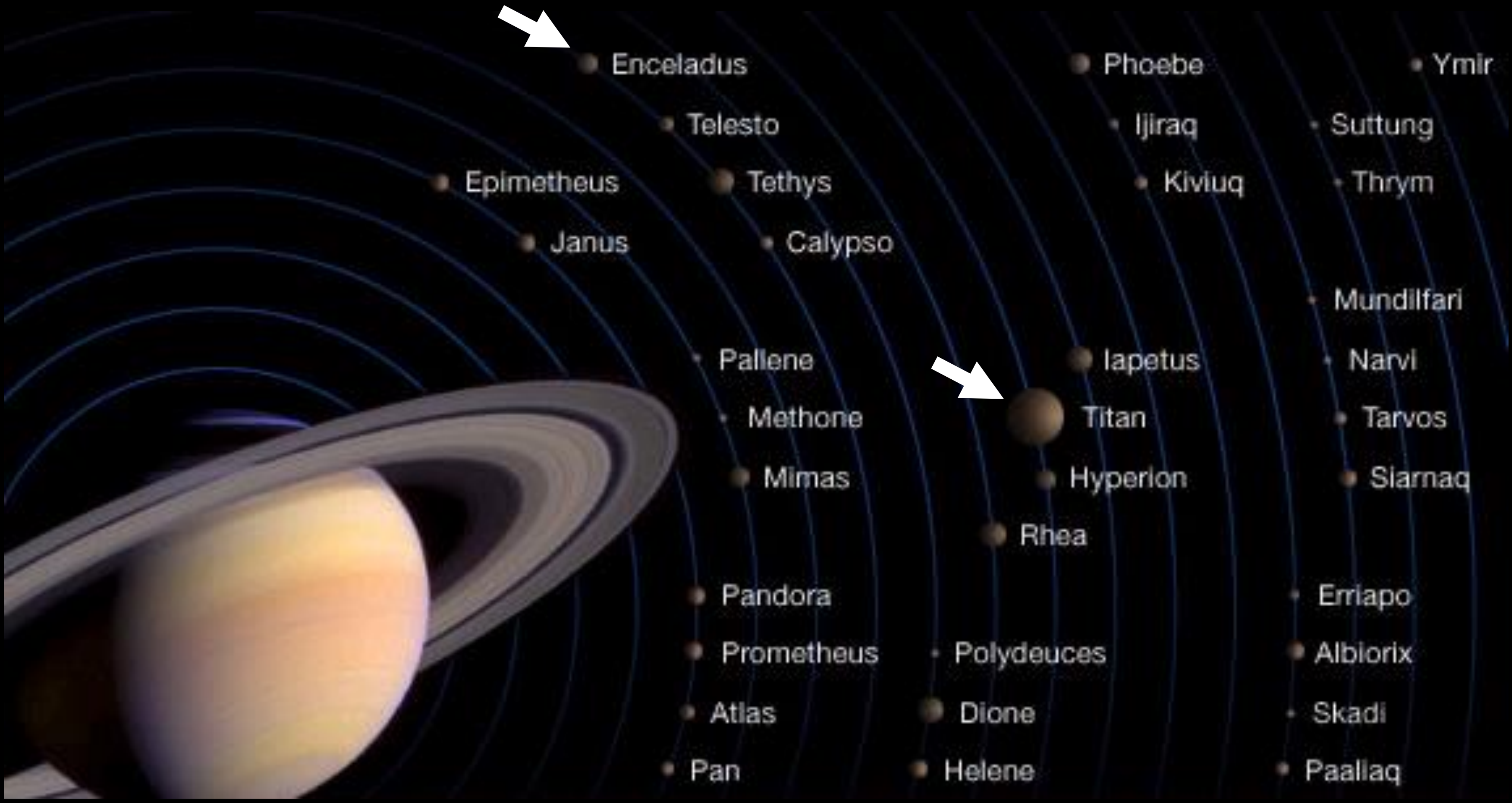
6 to 44 minute
round-trip communication delay

Missions to Jupiter’s Moons



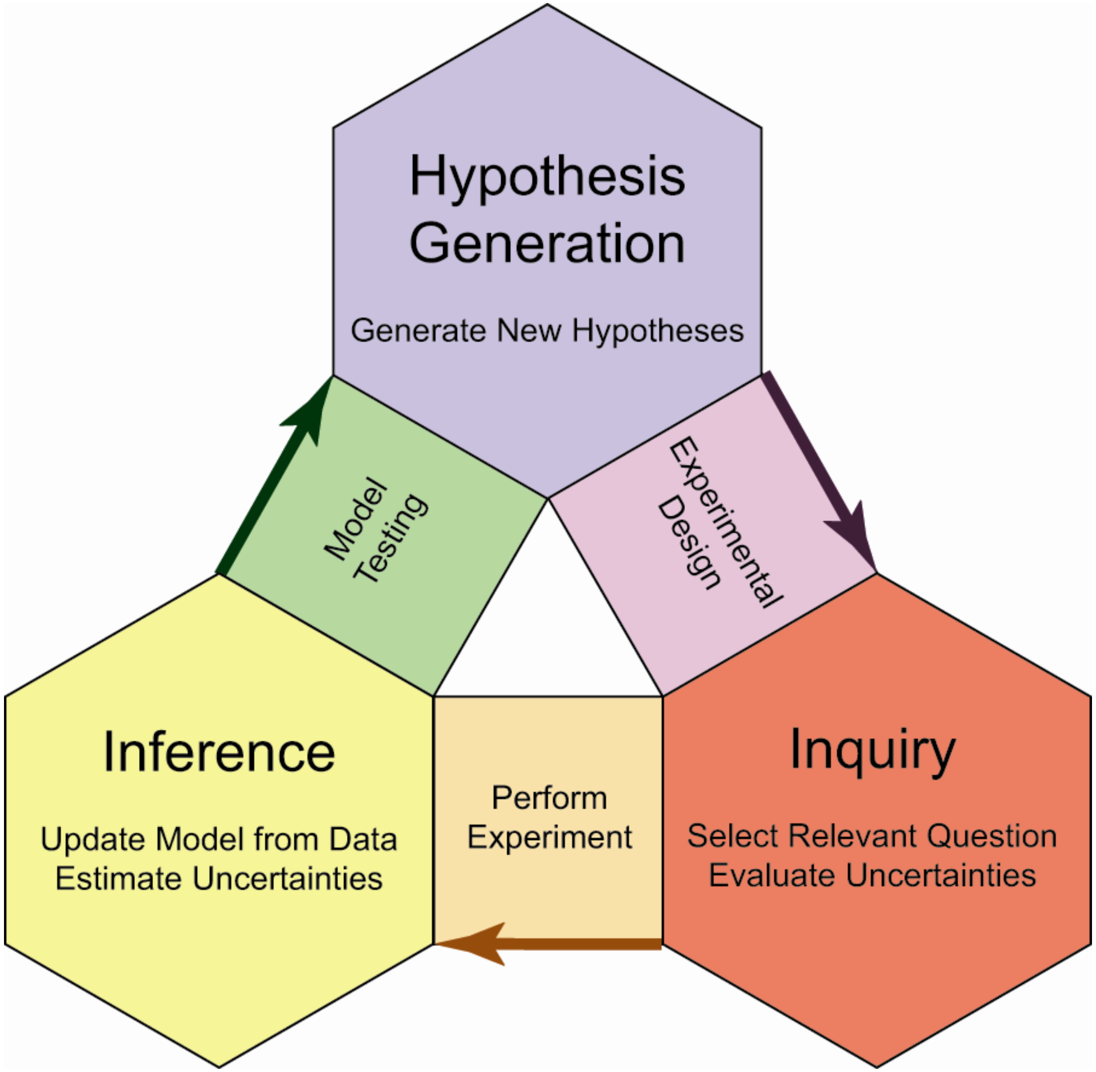
60 to 100 minute
round-trip communication delay

Missions to Saturn's Moons

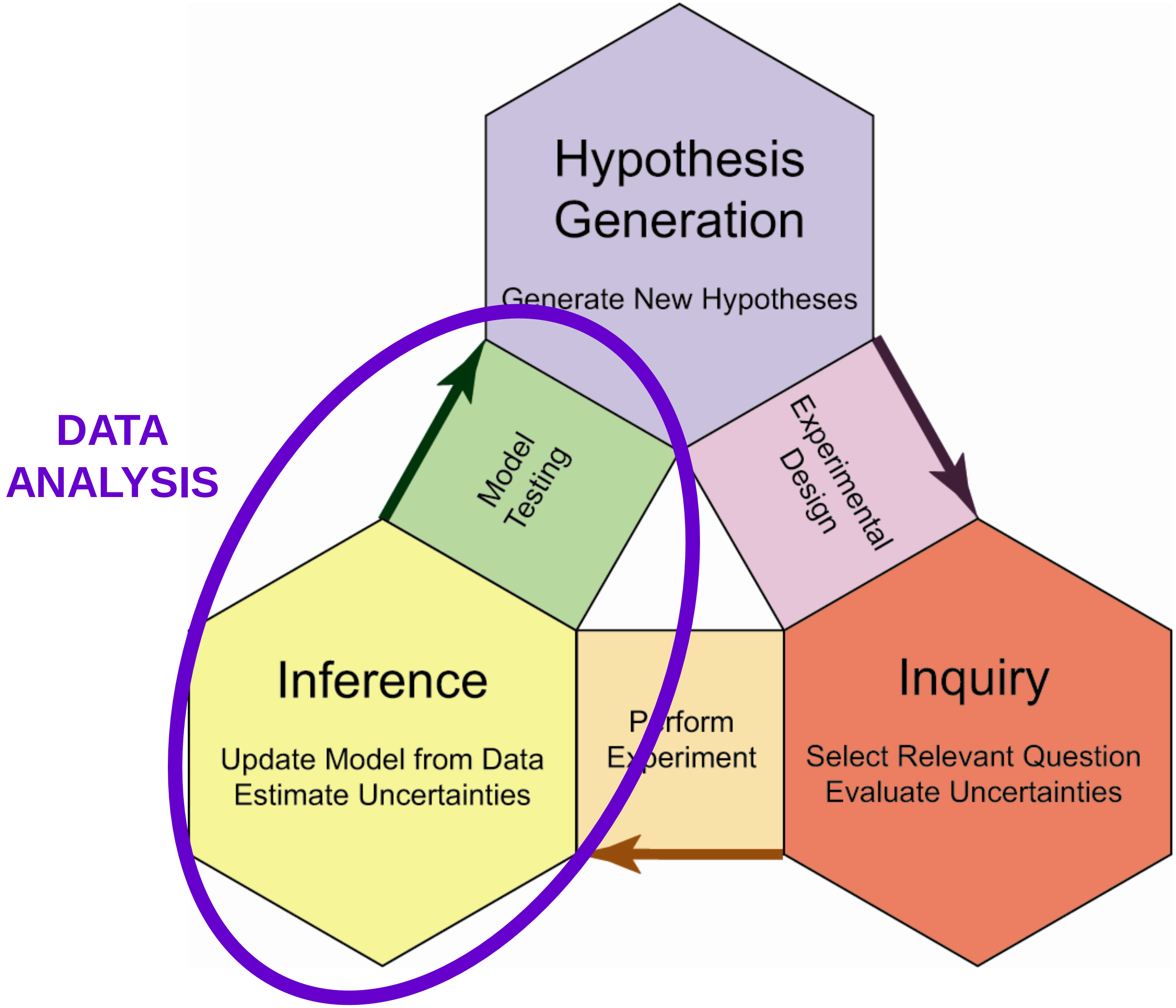


2.3 – 3 hour round-trip communication delay

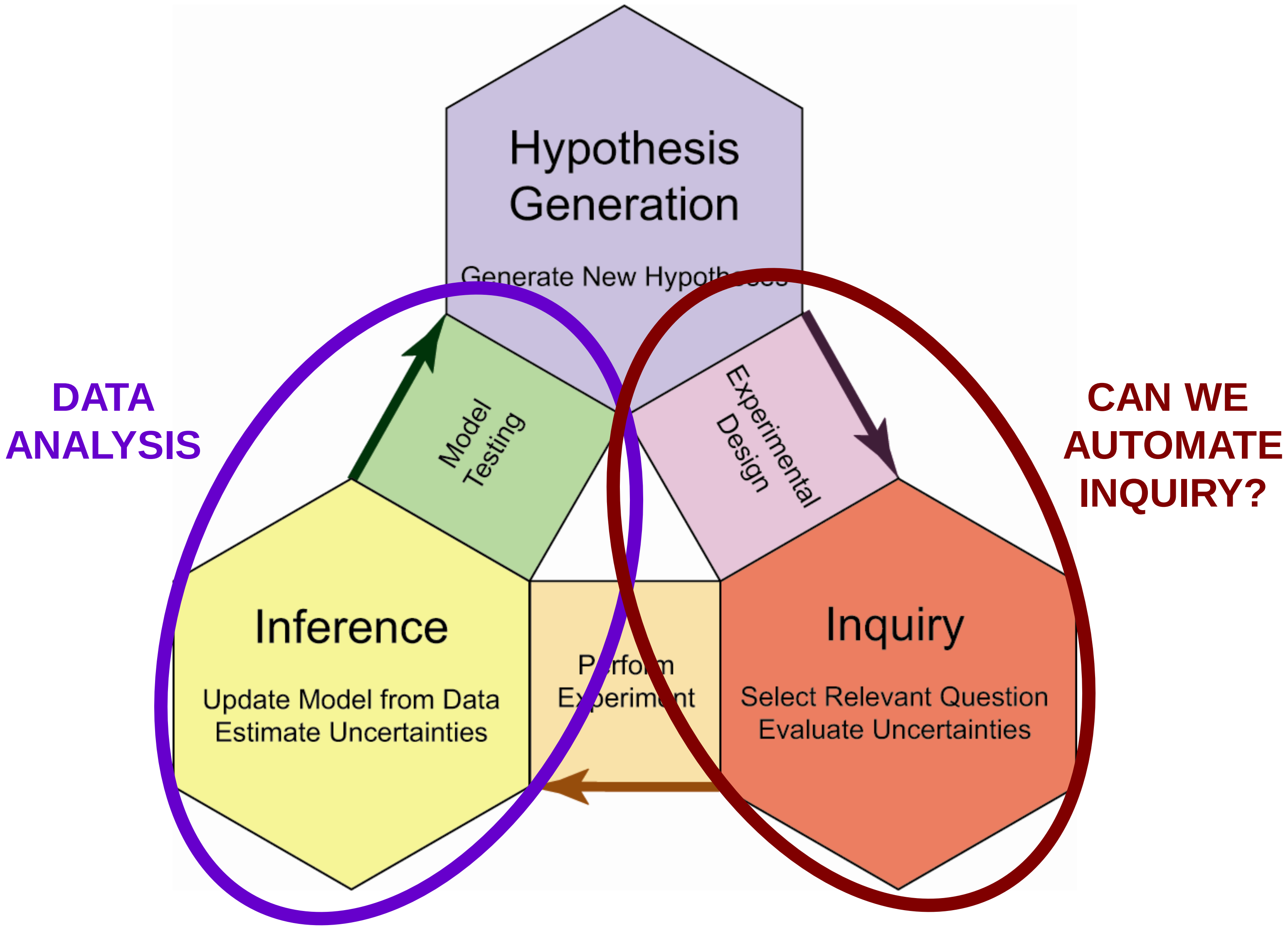
The Scientific Method



The Scientific Method



The Scientific Method



Describing the World

Partially Ordered Sets

a b c

Choosing a Piece of Fruit

apple

banana

cherry

State Space

apple

banana

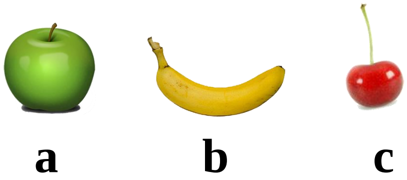
cherry

States describe Systems
Antichain

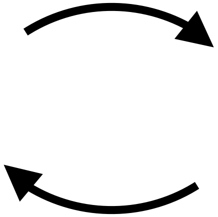
Exp and Log

$\bar{N}$

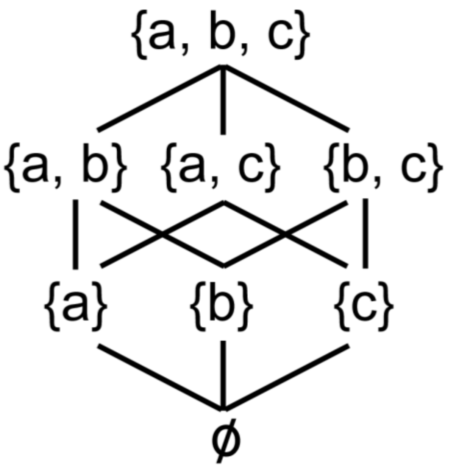
2^N



exp

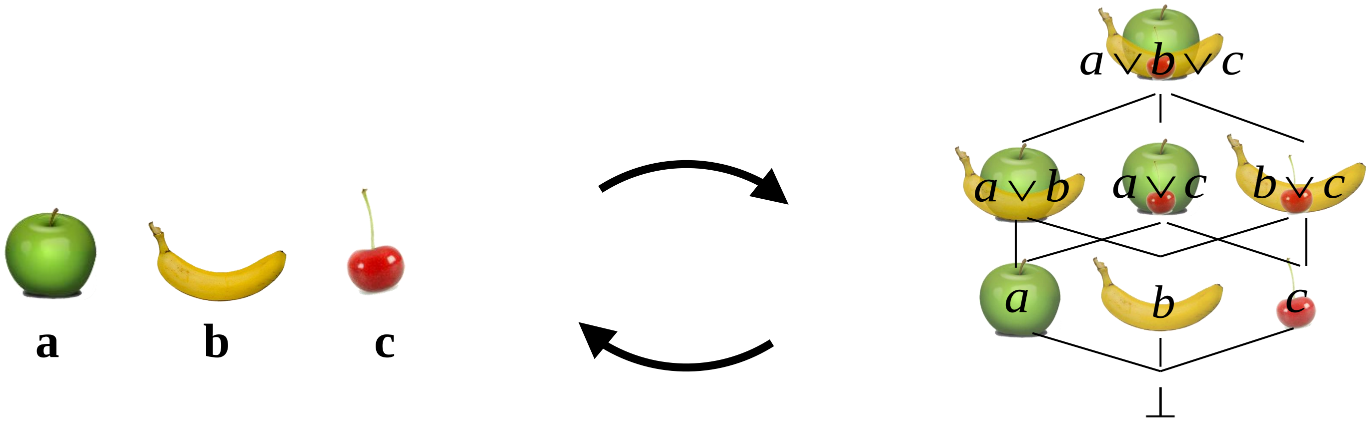


log



Exp and Log

$\overline{N}$ 2^N



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

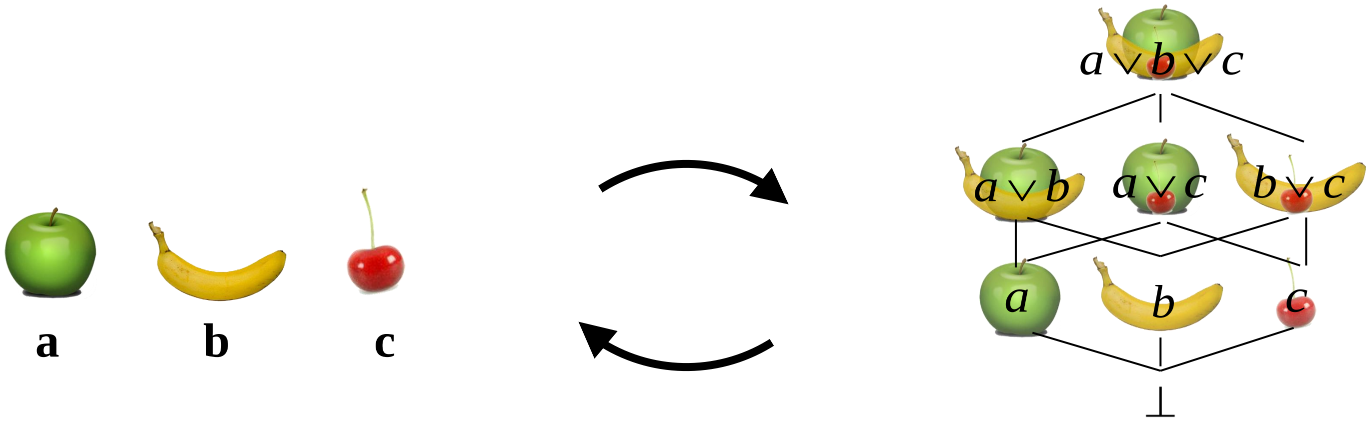
$$a \vee b \doteq \{\mathrm{a}, \mathrm{b}\}$$

$$\rightarrow \doteq \subseteq$$

Exp and Log

$\overline{N}$

2^N

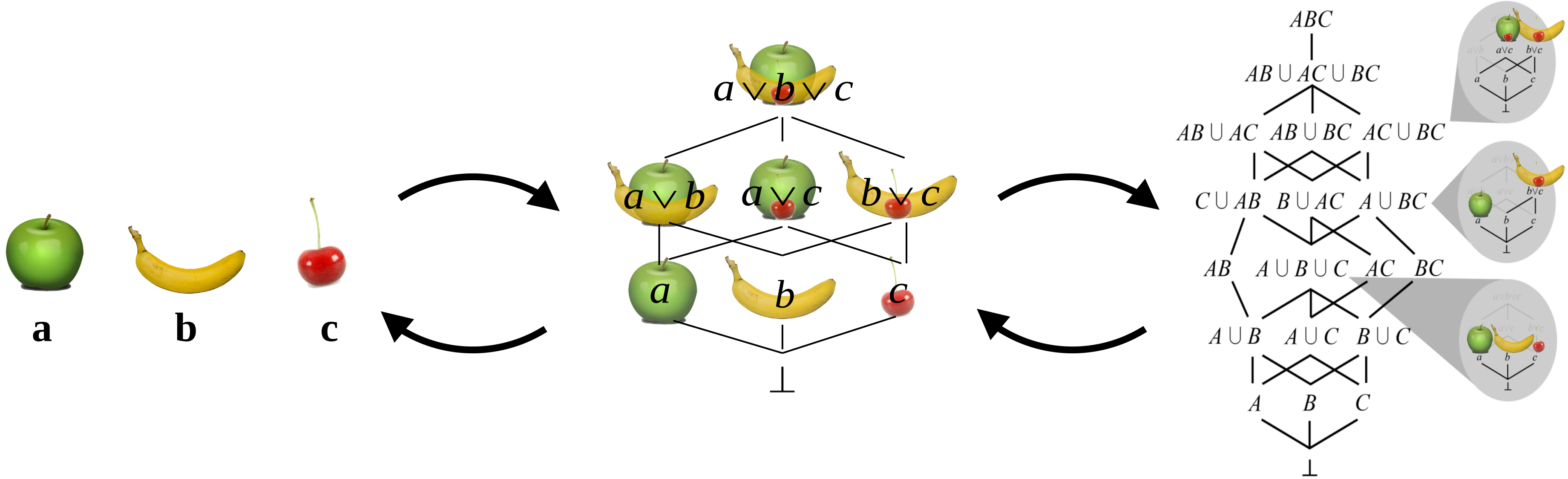


States

Statements
(sets of states)
(potential states)

Three Spaces

$\overline{N}$ 2^N $FD(N)$



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

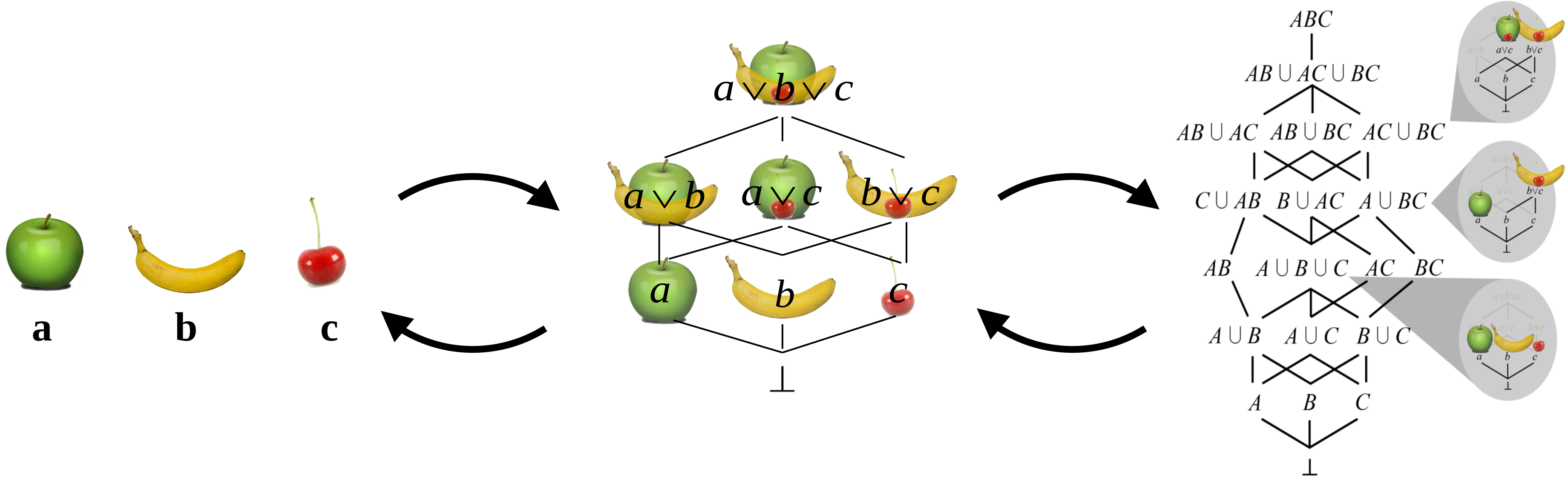
$$a \vee b \doteq \{\mathrm{a}, \mathrm{b}\}$$

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

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

Three Spaces

$\overline{N}$ 2^N $FD(N)$



States

Statements
(sets of states)
(potential states)

Questions
(sets of statements)
(potential statements)

State Space

apple

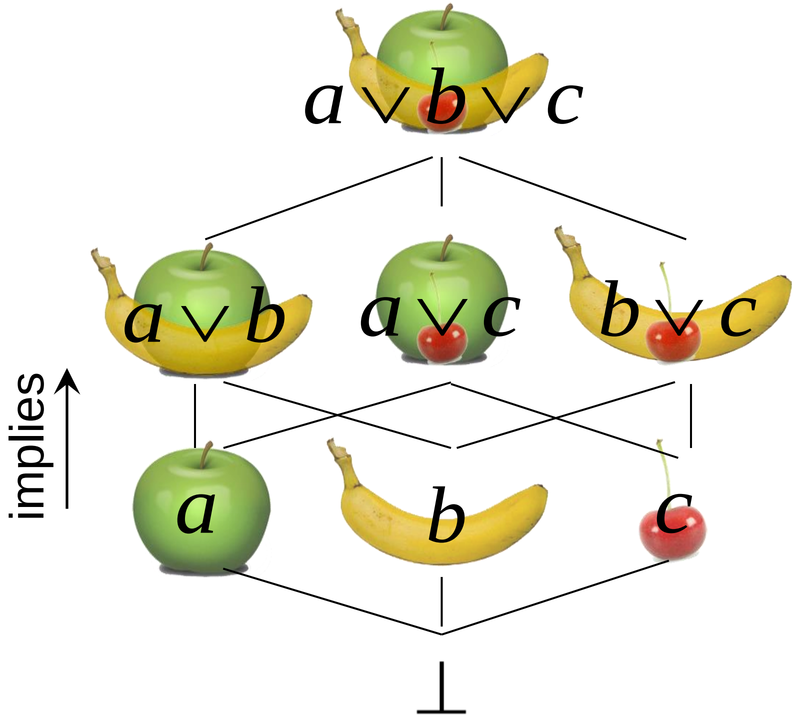
banana

cherry

States describe Systems

Antichain

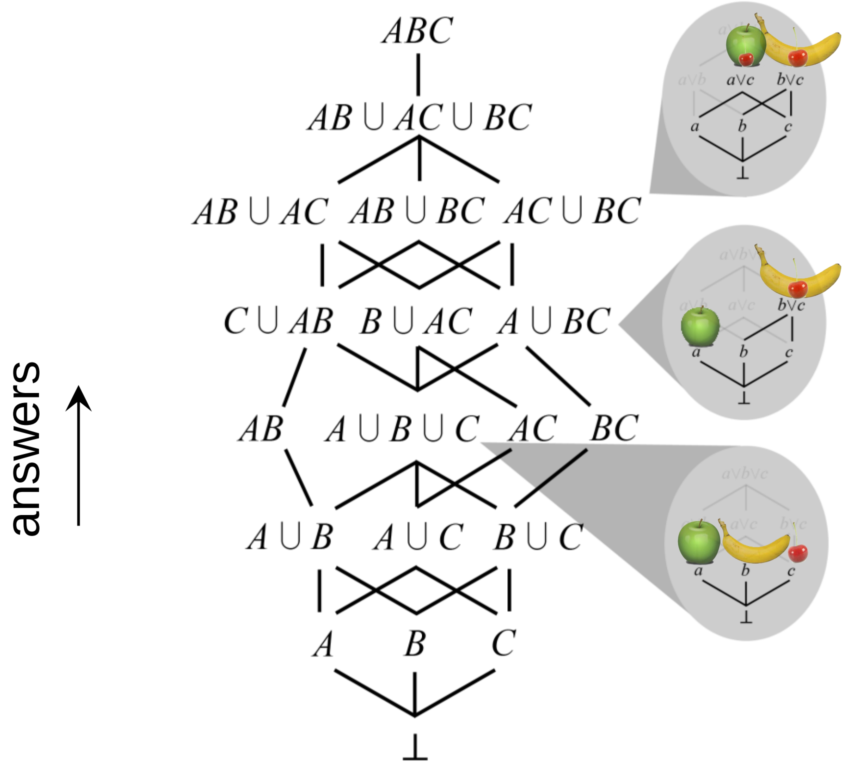
Hypothesis Space



Statements are sets of States

Boolean Lattice

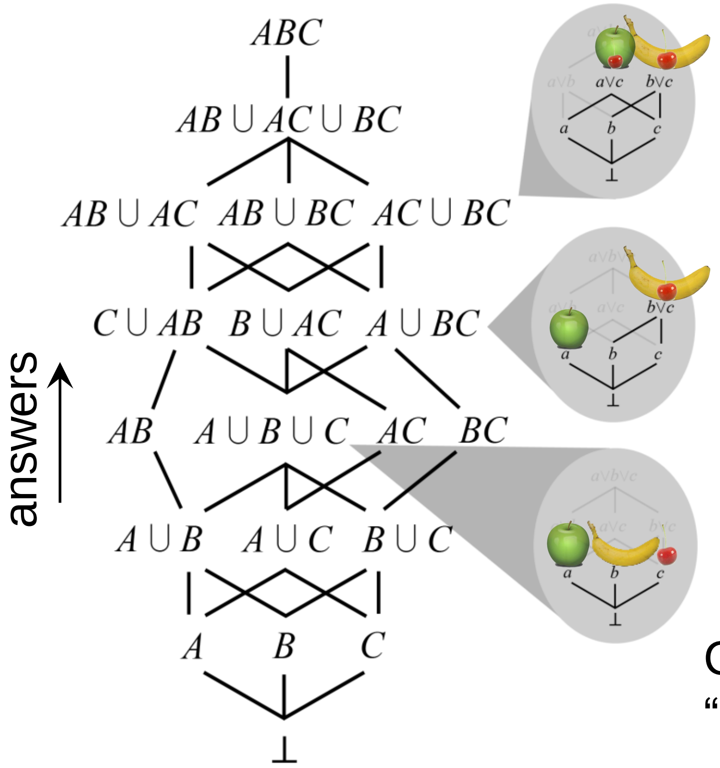
Inquiry Space



Questions are sets of Statements

Free Distributive Lattice

Relevance



“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?”

Relevance Decreases

The 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!" }

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

Some Questions Answer Others

Now consider the binary question

B = “Is it 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 of I is exhaustive, $\sim a = b \vee c$

Ordering Questions

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

Valuations
on
Lattices

Valuations

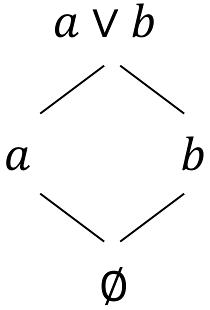
Valuations are functions
that take lattice elements to real numbers

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

Valuations

Valuations are functions
that take lattice elements to real numbers

Valuation: $v : x \in L \quad \rightarrow \quad \mathbb{R}$



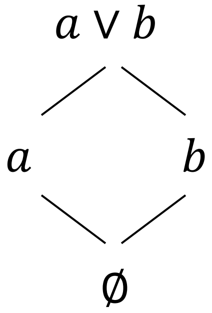
How do we ensure that the valuation assignments
are consistent with the lattice structure?

Local Consistency

Any general rule must hold for special cases.

Look at special cases to constrain general rule.

We enforce local consistency.



$$v(a \vee b) \leftrightarrow v(a) \text{ and } v(b)$$

This implies that:

$$v(a \vee b) = S[v(a), v(b)]$$

Associativity of Join V

Write the same element two different ways

$$a \vee (b \vee c) \quad = \quad (a \vee b) \vee c$$

This implies that:

$$S[v(a), S[v(b), v(c)]] \quad = \quad S[S[v(a), v(b)], v(c)]$$

Associativity of Join V

Write the same element two different ways

$$a \vee (b \vee c) \quad = \quad (a \vee b) \vee c$$

This implies that:

$$S[v(a), S[v(b), v(c)]] \quad = \quad S[S[v(a), v(b)], v(c)]$$

The general solution (Aczel) is:

$$F(S[v(a), v(b)]) \quad = \quad F(v(a)) + F(v(b))$$

$$m(a \vee b) \quad = \quad m(a) + m(b)$$

DERIVATION OF MEASURE THEORY!

Sum Rule

This result is known more generally as the SUM RULE

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

Context and Bi-Valuations

Bi-Valuation: $w : x, y \in L \quad \rightarrow \quad \mathbb{R}$

Bi-Valuation

Valuation

$$w(x\,|\,y) \longrightarrow v_y(x) \longrightarrow v(x)$$

Context y
is explicit

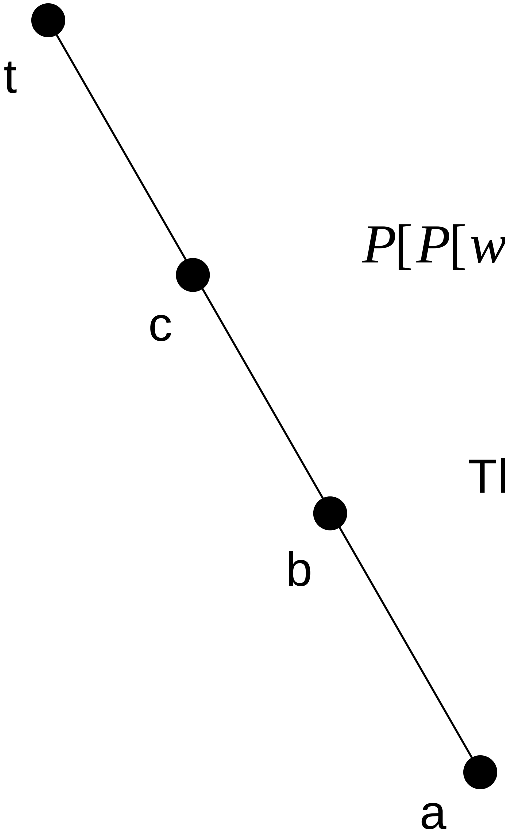
Measure of x
with respect to
Context y

Context y
is implicit

Bi-valuations generalize lattice inclusion to
degrees of inclusion.

The bi-valuation inherits meaning from the ordering relation!

Associativity of Context



$$w(a|t) = P[w(a|c), w(c|t)]$$

$$w(a|t) = P[w(a|b), w(b|t)]$$

$$P[P[w(a|b), w(b|c)], w(c|t)] = P[w(a|b), P[w(b|c), w(c|t)]]$$

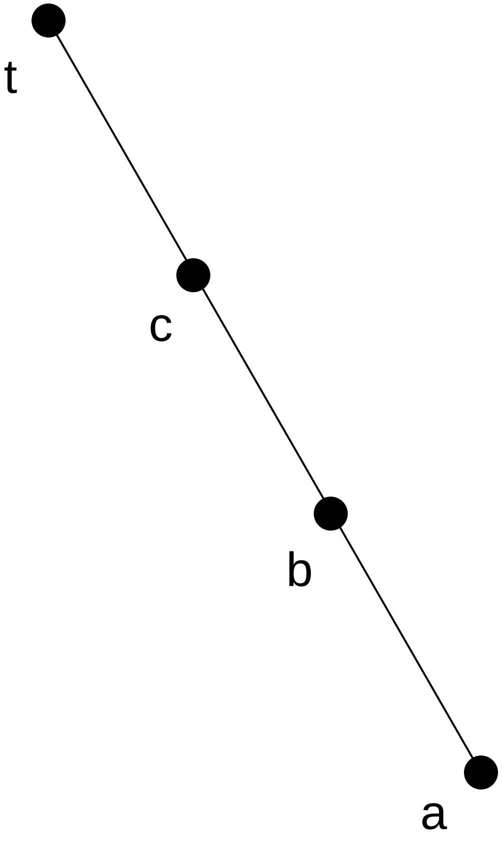
The Result:

$$G(P[w(a|c), w(c|t)]) = G(w(a|c)) + G(w(c|t))$$

$$m(a|t) = m(a|c)\, m(c|t)$$

Product Rule!

Product Rule and Context



$$m(a\,|\,t) \quad = \quad m(a\,|\,c)\,m(c\,|\,t)$$

Ratios of Measures

$$m(a\,|\,c) \quad = \quad \frac{m(a\,|\,t)}{m(c\,|\,t)}$$

In General: Two Product Rules

$$m(a \wedge c\,|\,t) \quad = \quad m(a\,|\,c \wedge t)\,m(c\,|\,t)$$

$$m(a\,|\,c \vee t) \quad = \quad m(a\,|\,c)\,m(a \vee c\,|\,t)$$

Commutativity

Commutativity $x \wedge y = x \wedge y$

leads to a **Bayes Theorem…**

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

Note that Bayes Theorem involves a change of context.

Valuations are not sufficient… need bi-valuations.

Inclusion-Exclusion (The Sum Rule)

$$w(x \vee y \mid t) \;=\; w(x \mid t) + w(y \mid t) - w(x \wedge y \mid t)$$

The Sum Rule for Lattices

Inclusion-Exclusion (The Sum Rule)

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

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

The Sum Rule for Probability

Inclusion-Exclusion (The Sum Rule)

$$w(x \vee y \mid t) \; = \; w(x \mid t) \; + \; w(y \mid t) \; - \; w(x \wedge y \mid t)$$

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

Definition of Mutual Information

Inclusion-Exclusion (The Sum Rule)

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

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

Polya’s Min-Max Rule for Integers

Inclusion-Exclusion (The Sum Rule)

$$w(x \vee y \mid t) \;=\; w(x \mid t) \;+\; w(y \mid t) \;-\; w(x \wedge y \mid t)$$

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

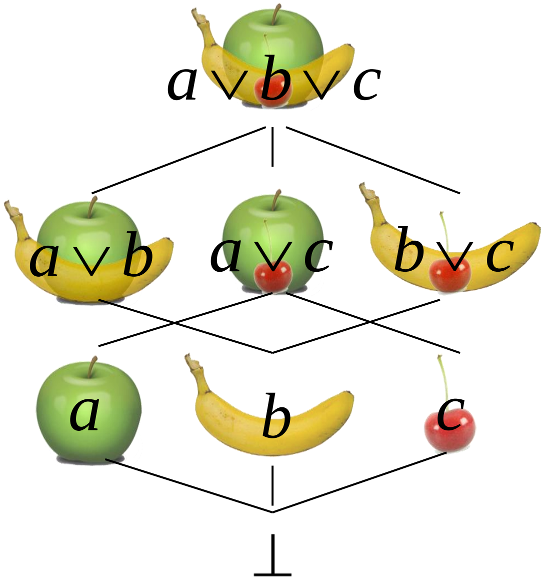
“Measuring Integers”, Knuth 2009

The Sum Rule derives from the Möbius function of the lattice,
And is related to its Zeta function

Probability

Probabilities are degrees of implication!

$$w(a \mid t) \equiv p(a \mid t)$$



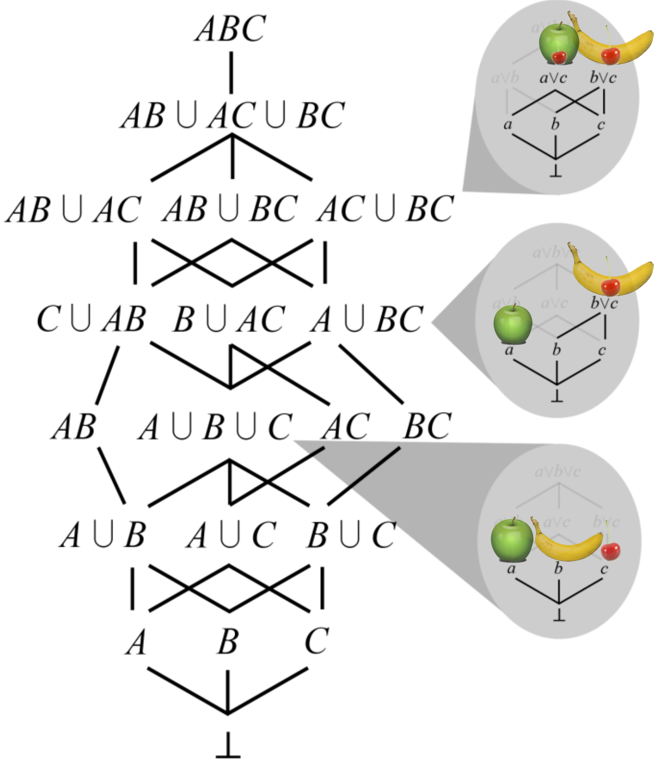
Constraint Equations!

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

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

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

Relevance



Relevance quantifies the degree to which one question answers another

$$d(I \mid A)$$

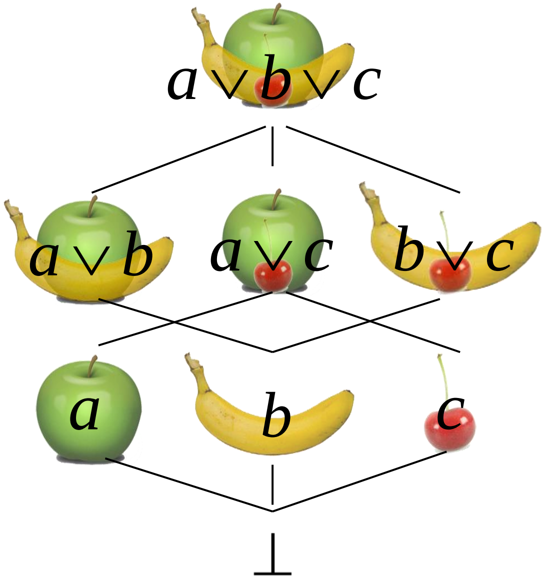
Constraint Equations

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

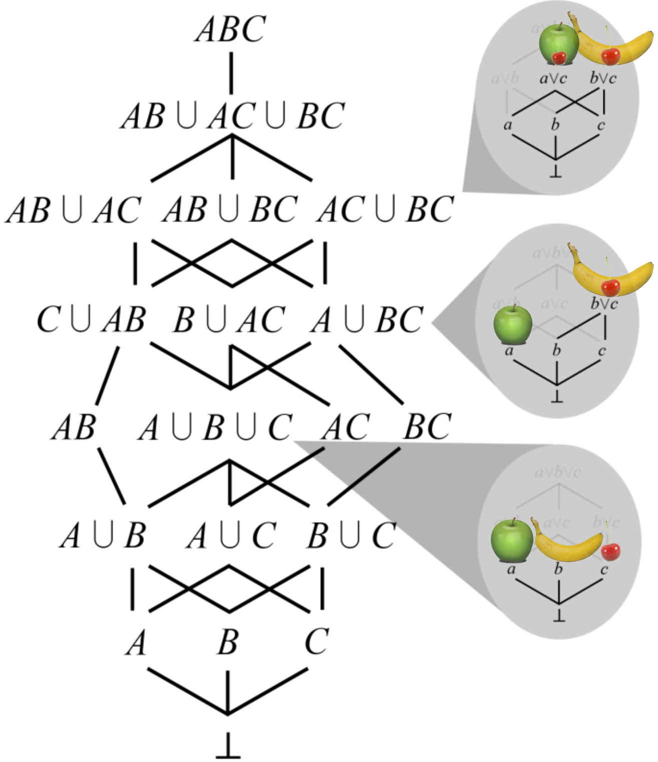
$$d(I \mid A \vee B) = d(I \mid A)\, d(A \vee I \mid B)$$

$$d(A \mid B) = \frac{d(I \mid B) d(B \mid A)}{d(I \mid A)}$$

Probability and Relevance

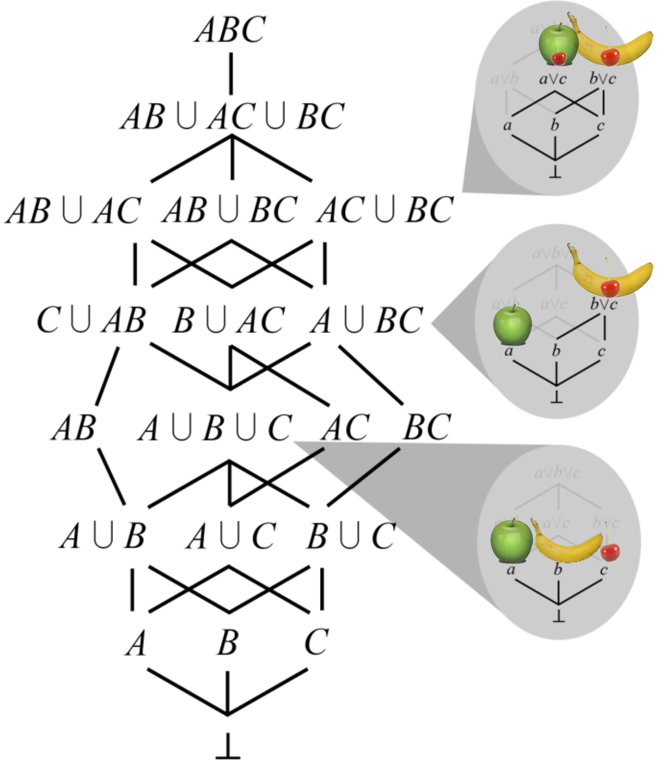


Relevance is a function of probability



The degree to which one question answers another must depend on the probabilities of the possible answers.

Relevance

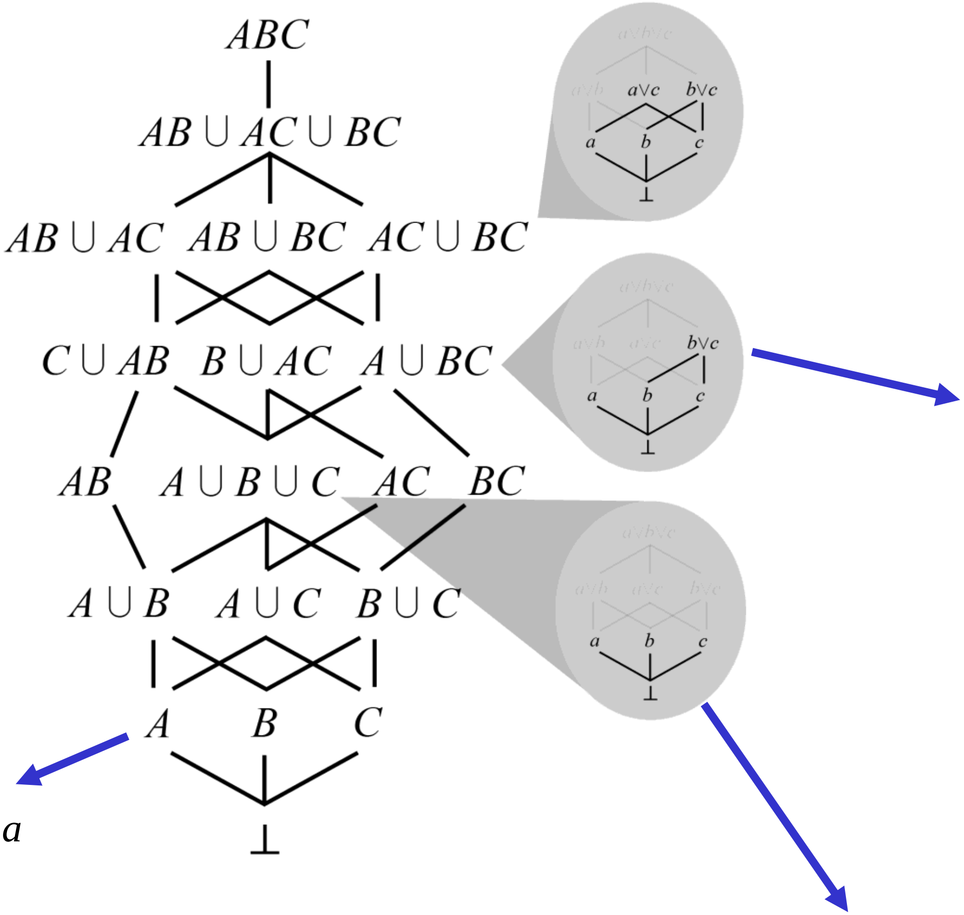


$$d(I \mid Q) = aH(Q) + b$$

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

Relevance and Entropy

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



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

$$-\,p_a \log_2 p_a$$

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

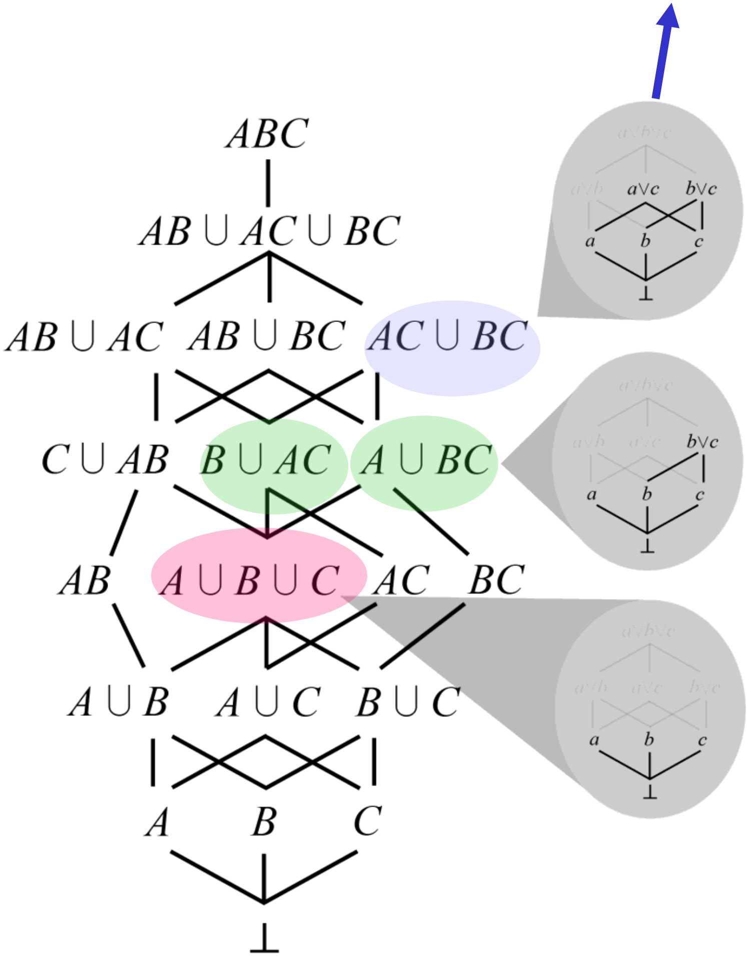
6 February 2009

Kevin H Knuth
CESS 2009

Higher-Order Informations

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

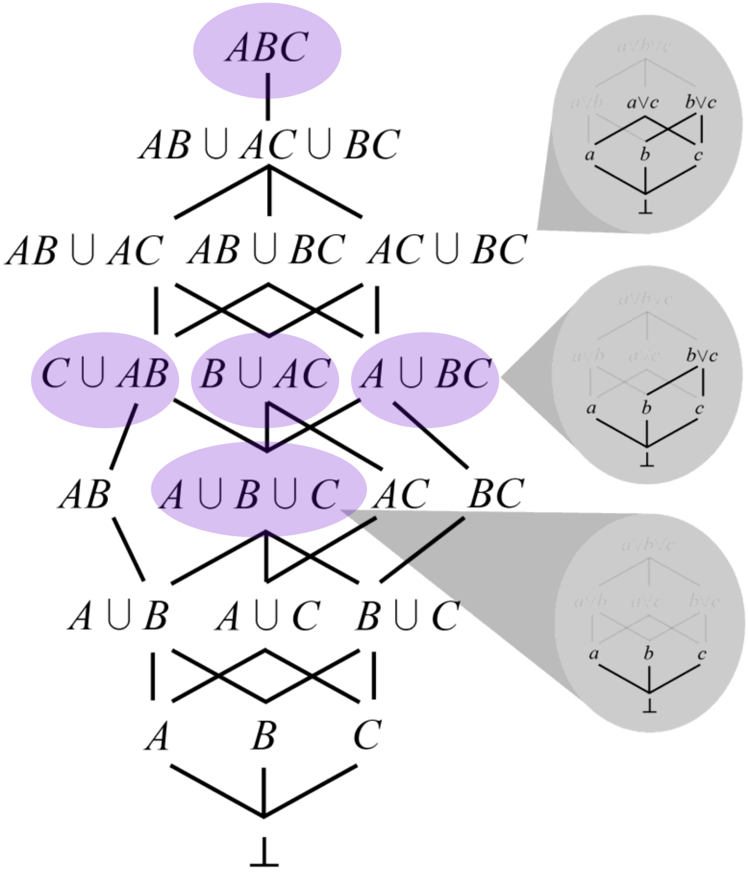
$$d(I \mid AC \cup BC) \sim I(B \cup AC; A \cup BC)$$



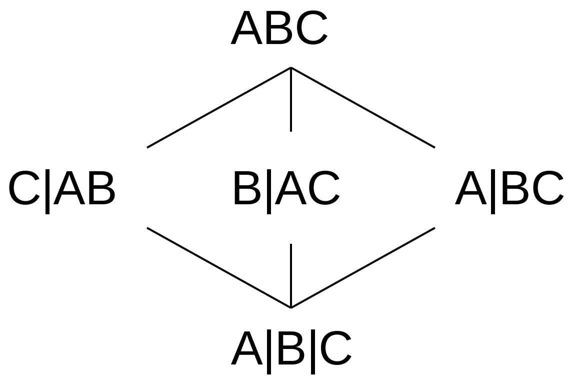
This relevance is related to the mutual information.

In this way one can obtain higher-order informations.

Partition Questions



Relevance is only a valid measure on the sublattice of questions isomorphic to partitions



EXAMPLE

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?

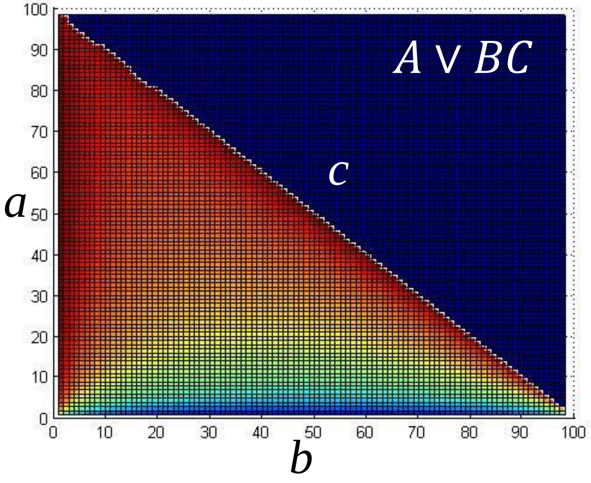
Is it or is it not a Banana?

Is it or is it not a Cherry?

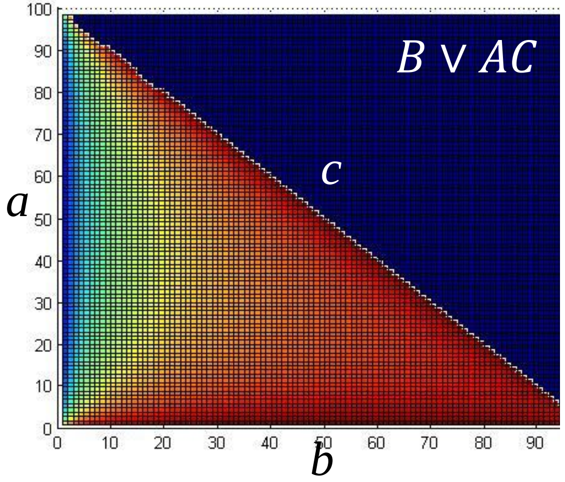
If you believe that there is a
75% chance that it is an Apple,
and a 10% chance that it is a Banana,
which question do you ask?

Relevance Depends on Probability

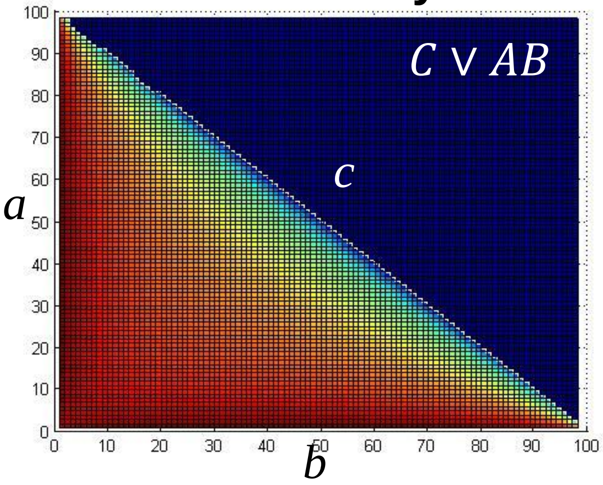
Is it an Apple?



Is it a Banana?



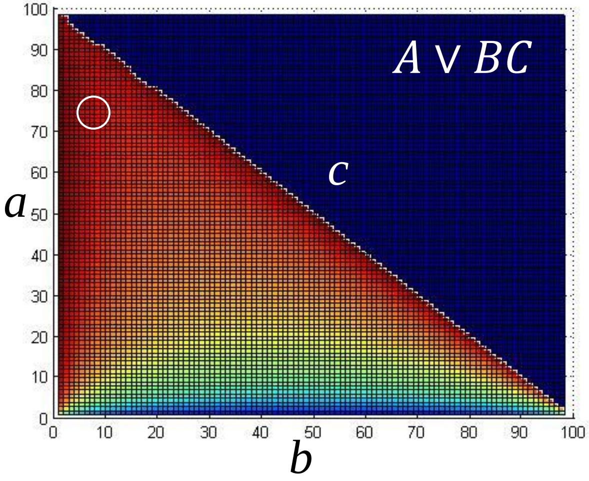
Is it a Cherry?



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?

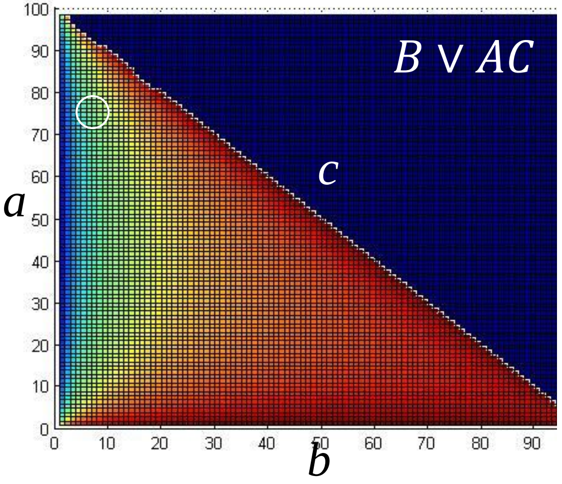
Relevance Depends on Probability

Is it an Apple?



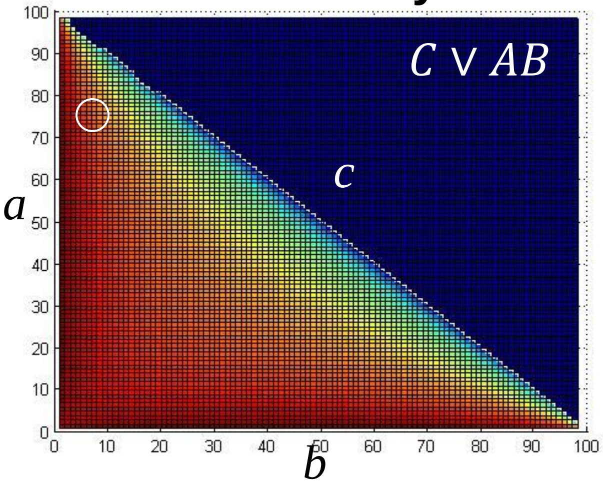
$$d(I \mid A \cup BC) \propto 0.5623$$

Is it a Banana?



$$d(I \mid B \cup AC) \propto 0.3250$$

Is it a Cherry?



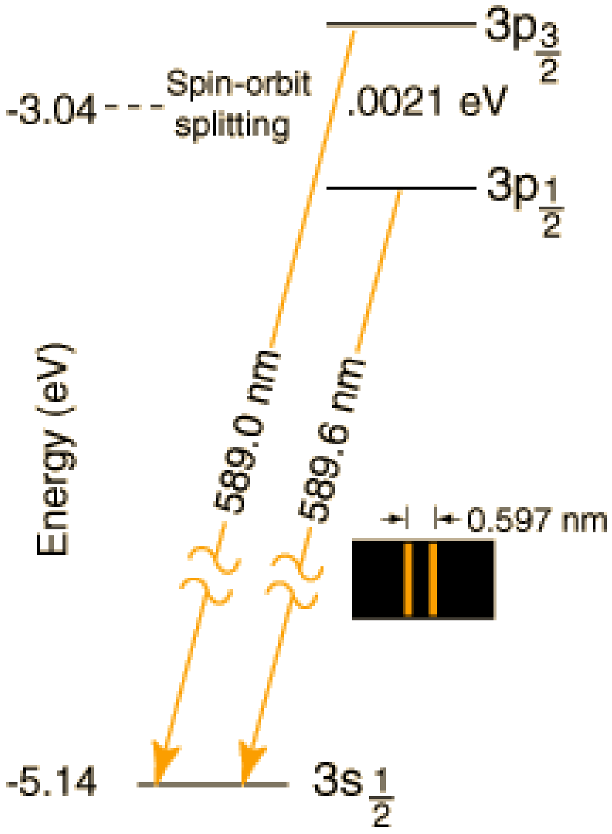
$$d(I \mid C \cup AB) \propto 0.4227$$

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?

EXPERIMENTAL DESIGN

Doppler Shift

PROBLEM:
Determine the relative radial velocity relative to a Sodium lamp. We can measure light intensities near the doublet at 589 nm and 589.6 nm



We can take ONE MEASUREMENT

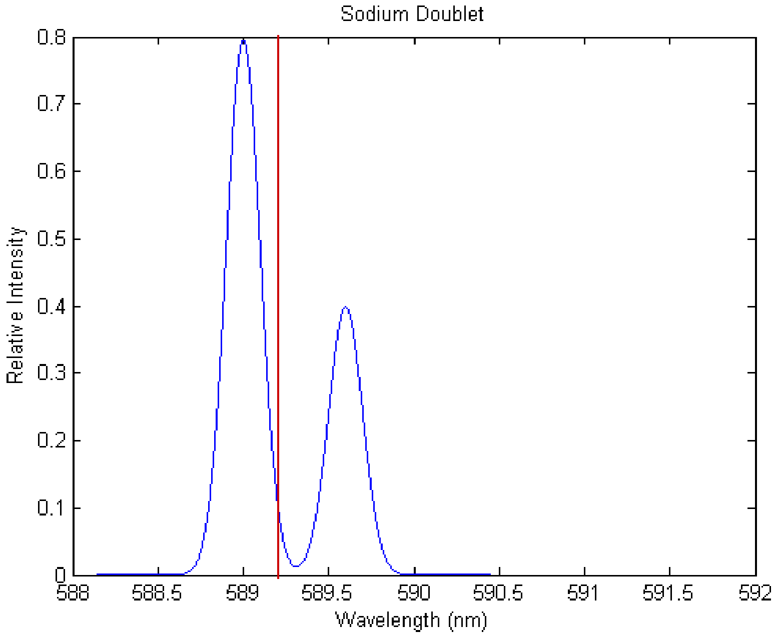
Which wavelength shall we examine?

Recall, we don’t know the Doppler shift!

What Can We Ask?

The question that can be asked is:

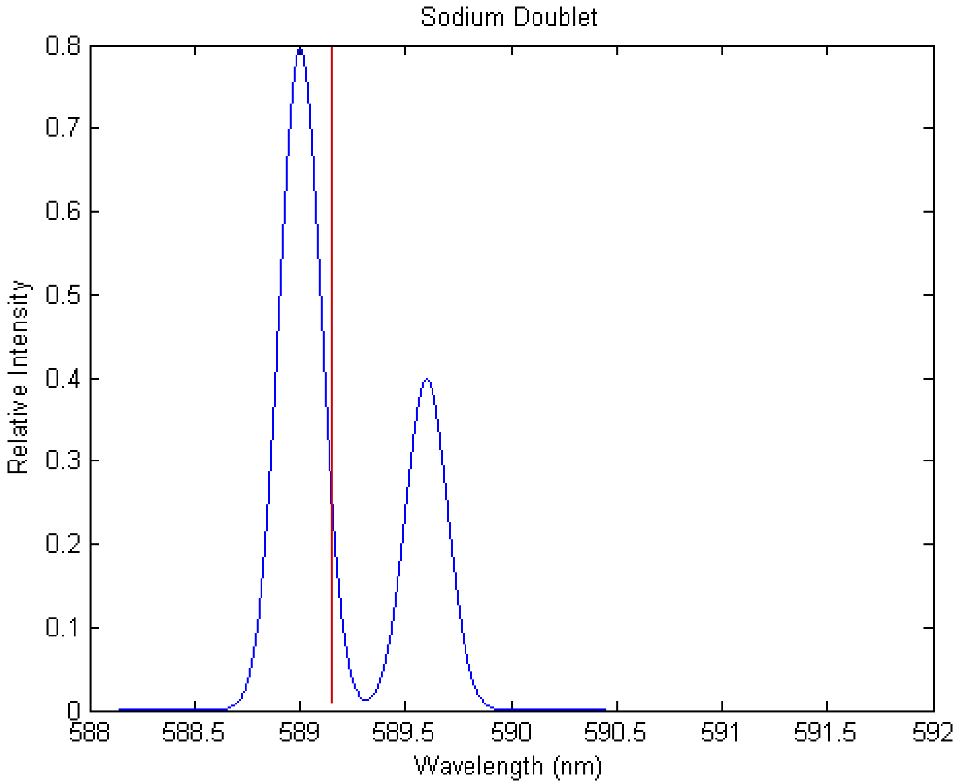
“What is the intensity at wavelength λ ?”



There are many questions to choose from, each corresponding to a different wavelength λ

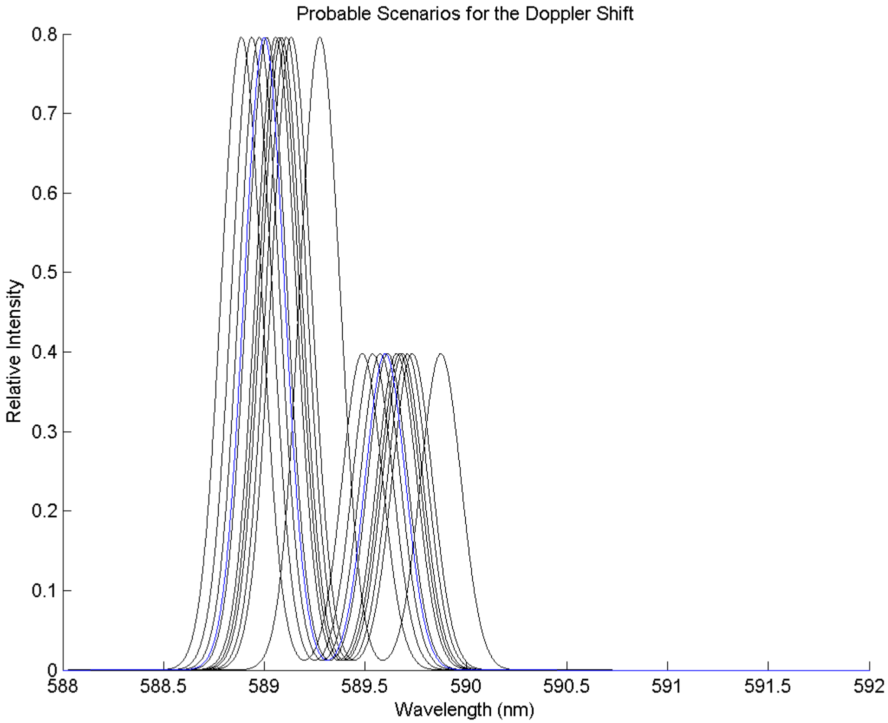
What are the Possible Answers?

Say that the intensity can be anywhere between 0 and 1.



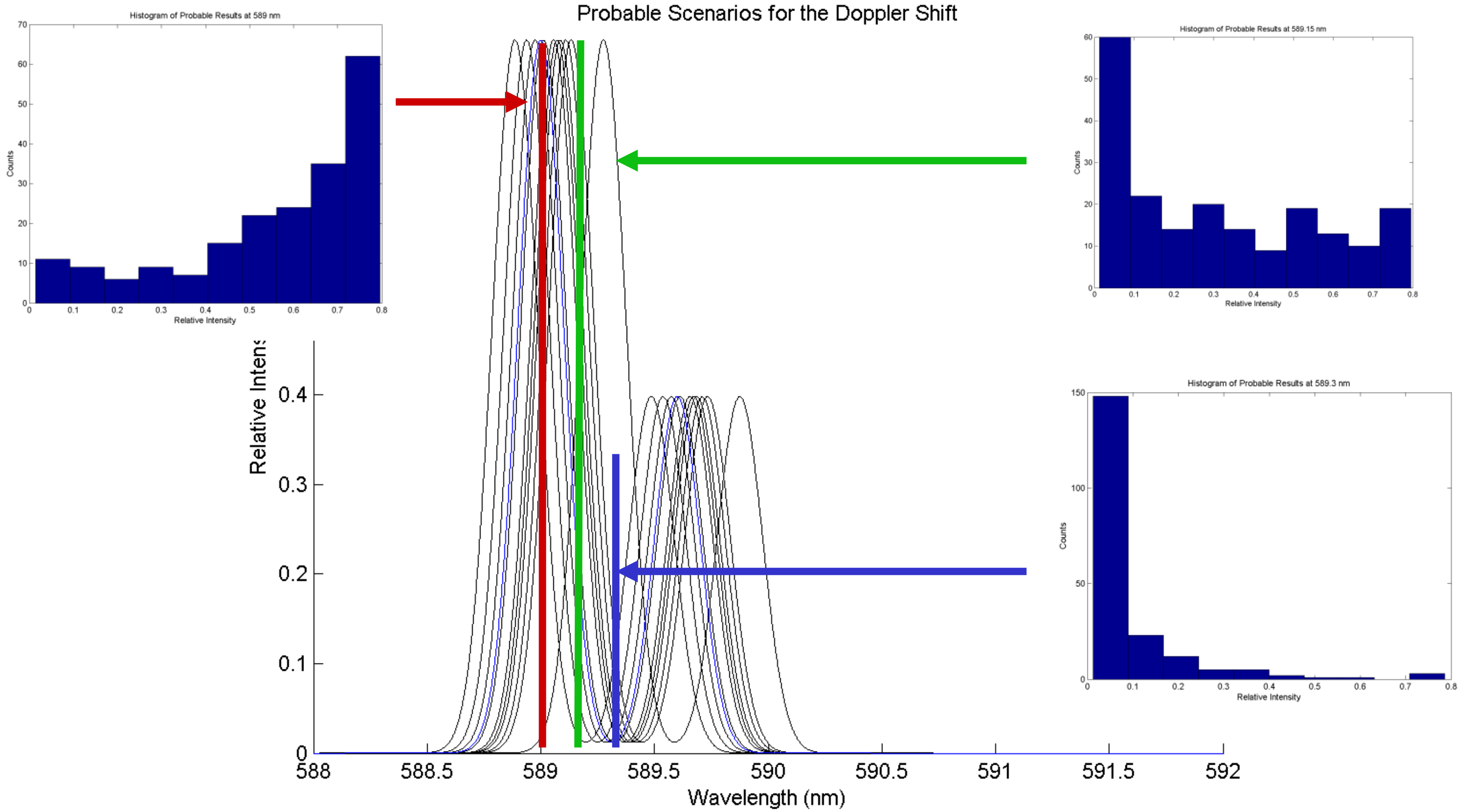
Given Possible Doppler Shifts…

Say we have information about the velocity.
The Doppler shift is such that the shift in wavelength has zero mean with a standard deviation of 0.1 nm.



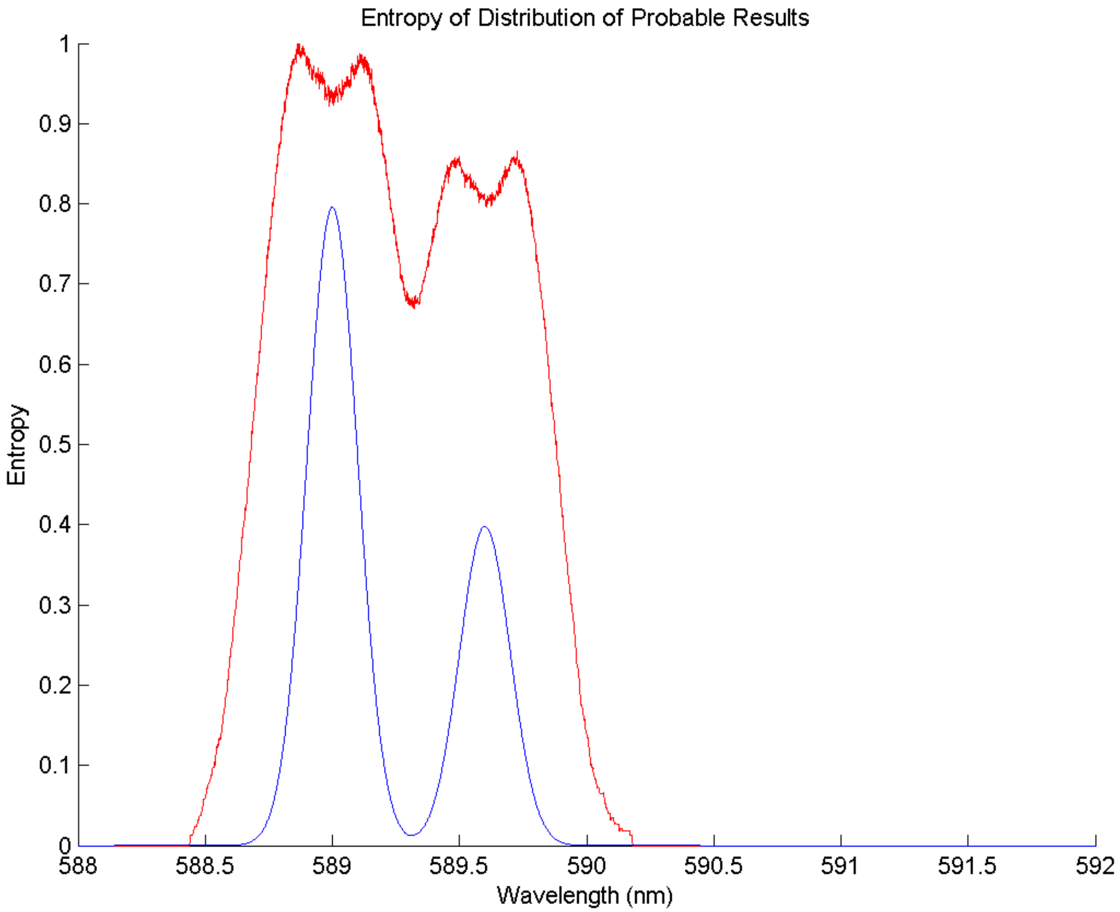
Probable Answers for Each Question

We now look at the set of probable answers for each question



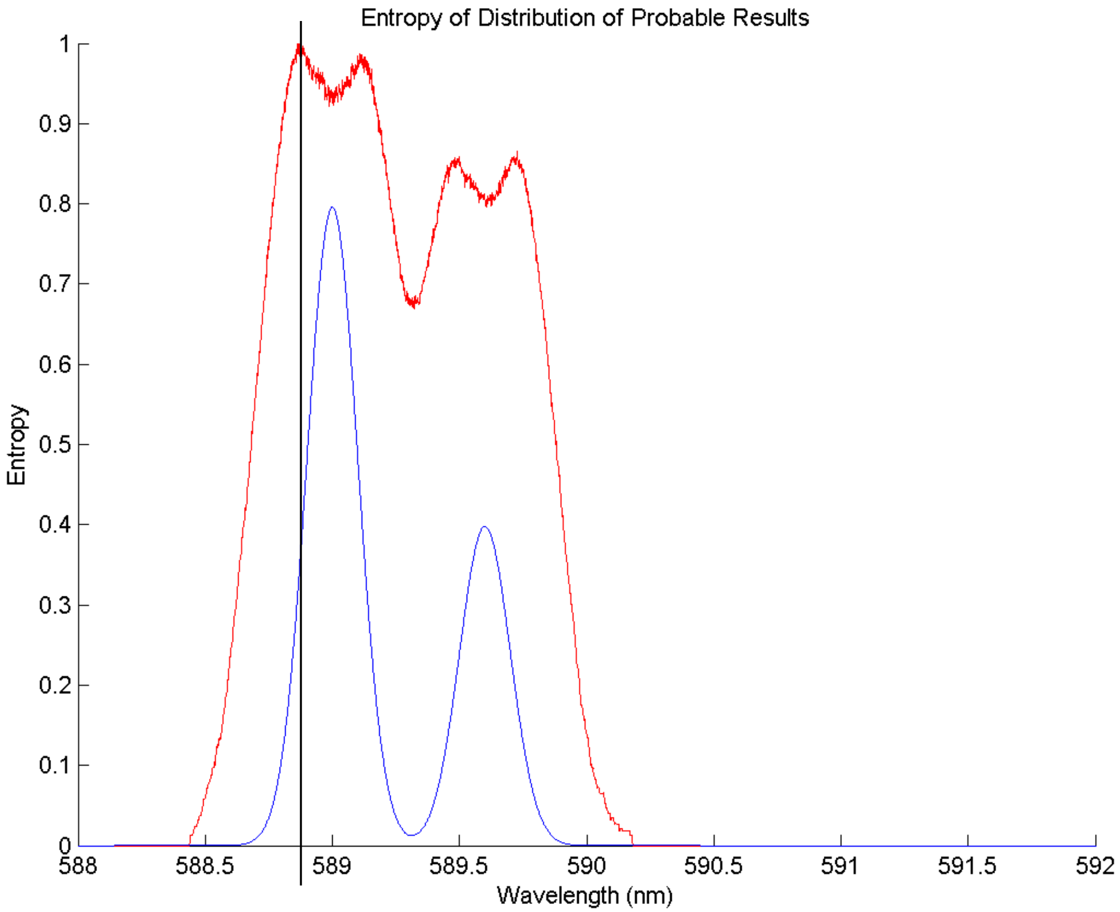
Entropy of Distribution of Probable Results

Red shows the entropy of the distribution of probable results.



Where to Measure???

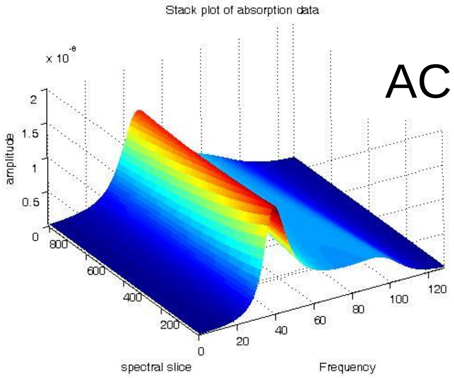
Measure where the entropy is highest!

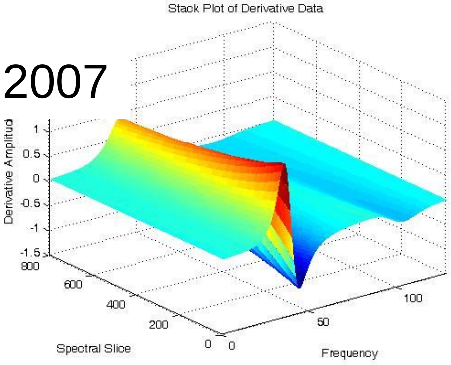


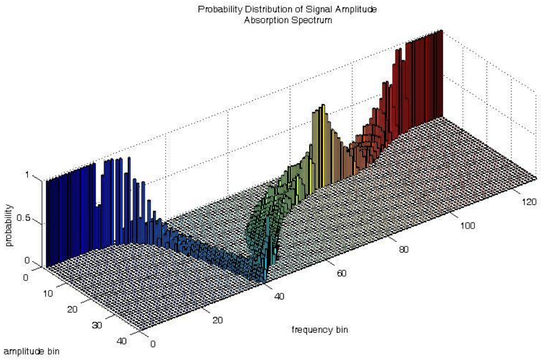
Professor Keith Earle

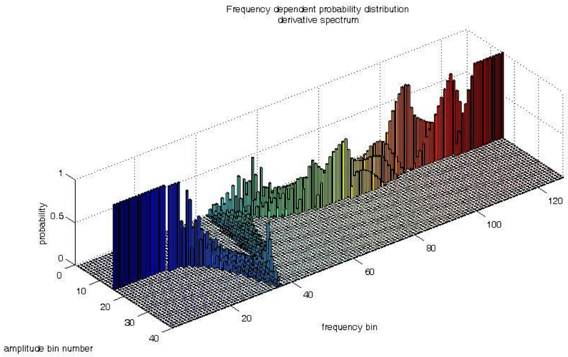
UAlbany (SUNY)

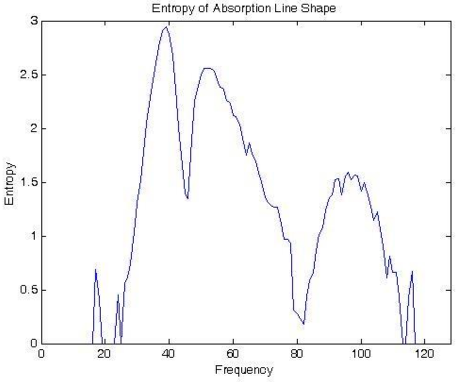
ACERT Simulation Workshop 2007

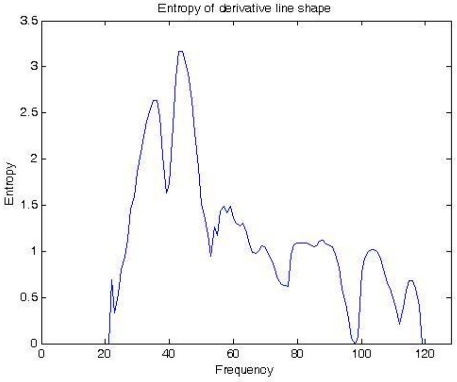










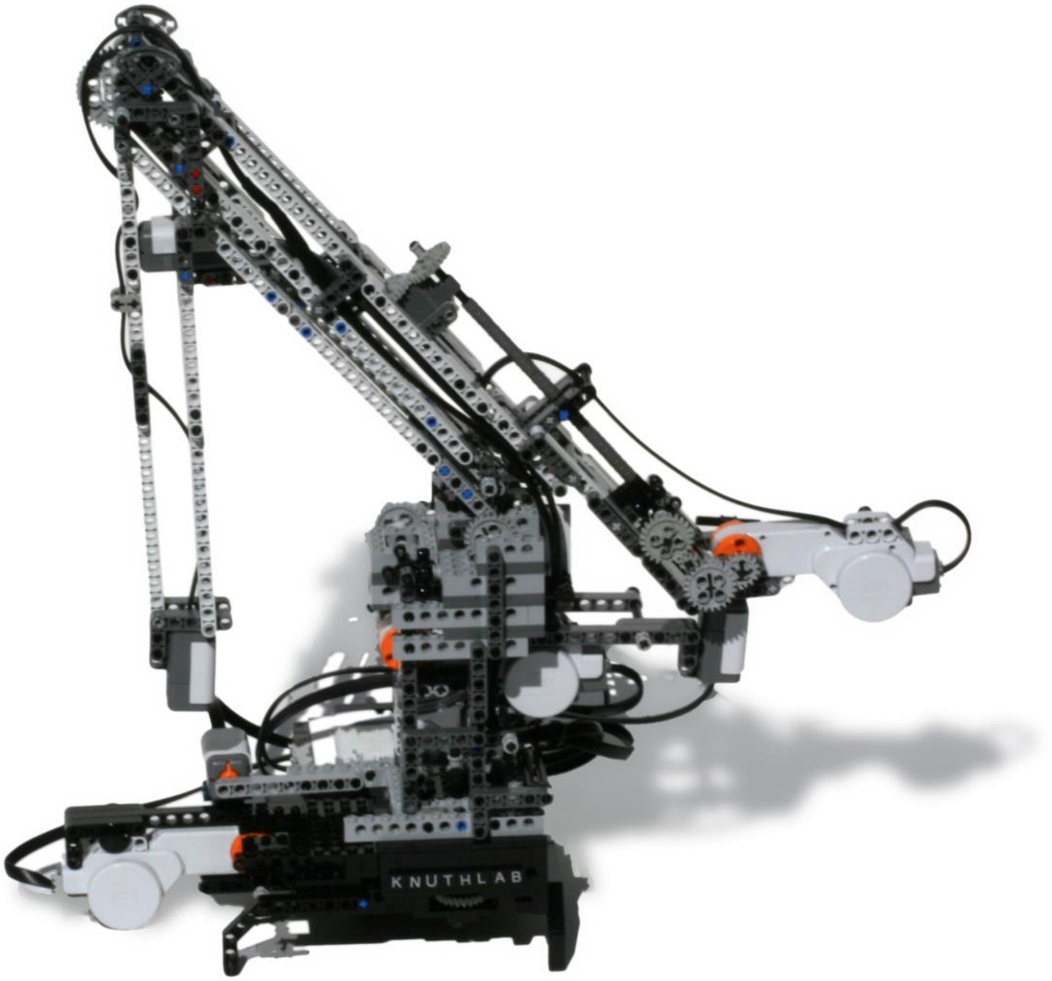


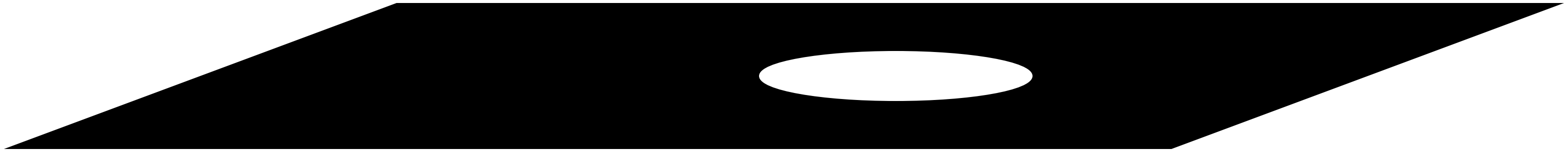
AUTOMATED INQUIRY

Robotic Scientists

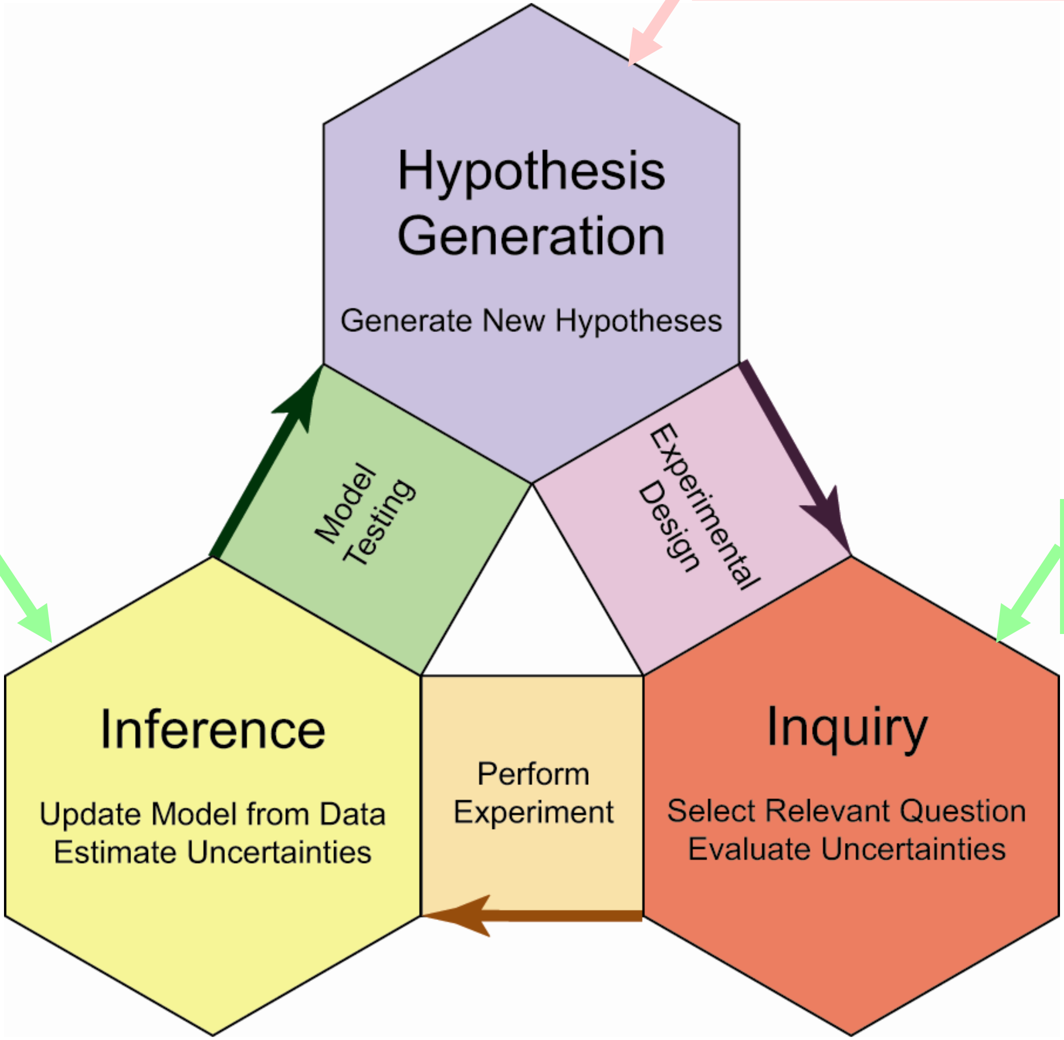
This robot is equipped with a light sensor.

It is to locate and characterize a white circle on a black playing field with as few measurements as possible.





Software Engines

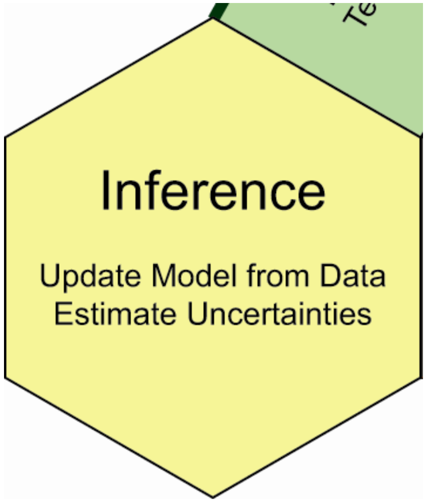


Preprogrammed

Implemented: Autonomous

Implemented: Autonomous

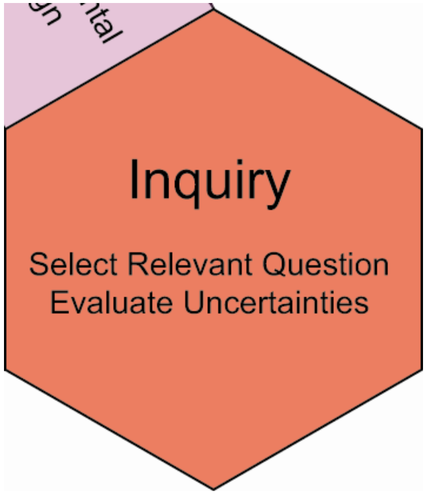
Inference Engine



Fully Bayesian Inference Engine

- Accommodates point spread function of light sensor
- Employs Nested Sampling (Skilling 2005) enabling automatic model selection
- Produces sample models from posterior probability

Inquiry Engine

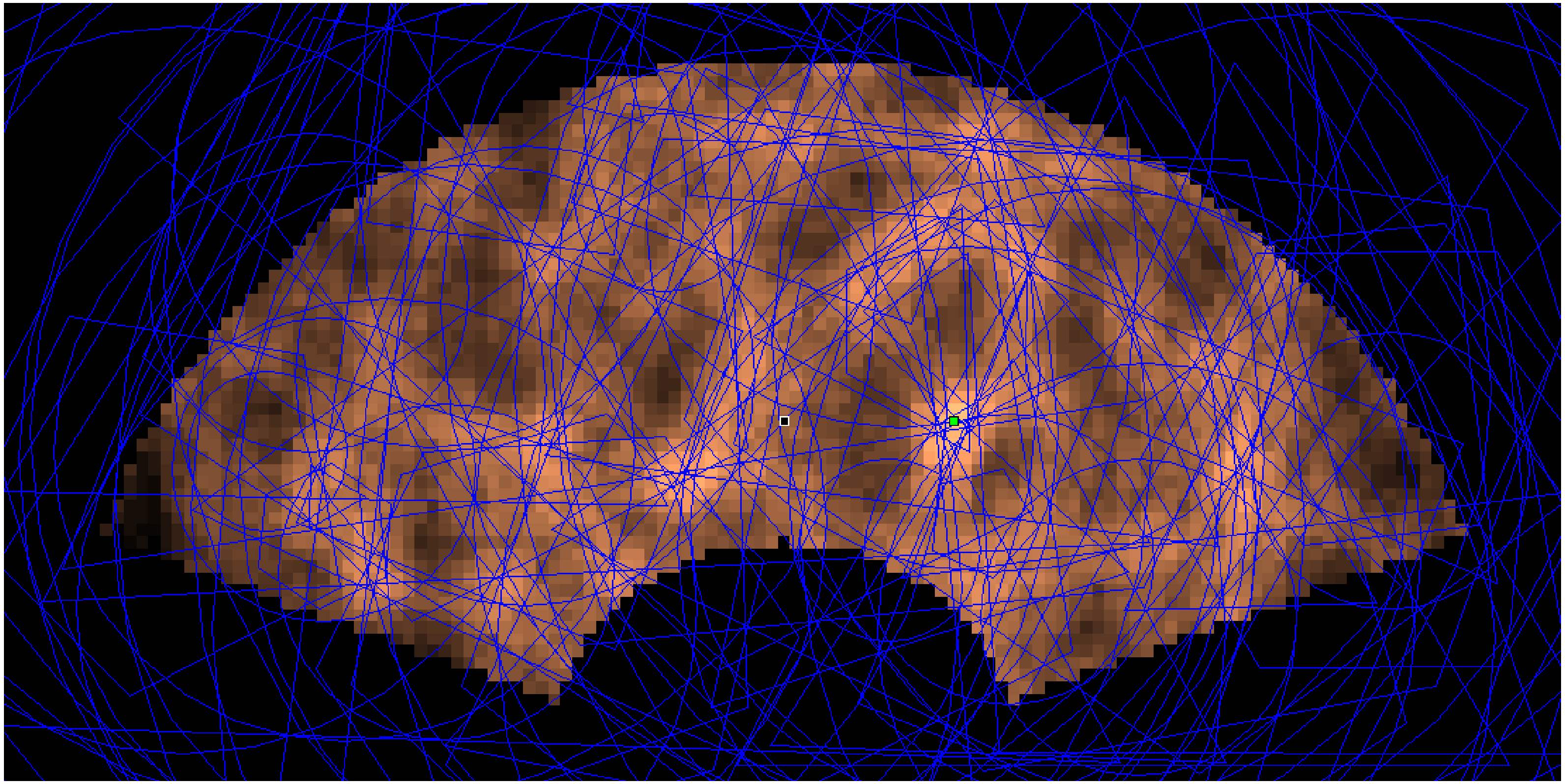


Autonomous Inquiry Engine

- Accommodates point spread function of light sensor
- Relies on samples provided by Inference Engine
- Rapid computation of entropy of distribution of measurements predicted by the sampled models

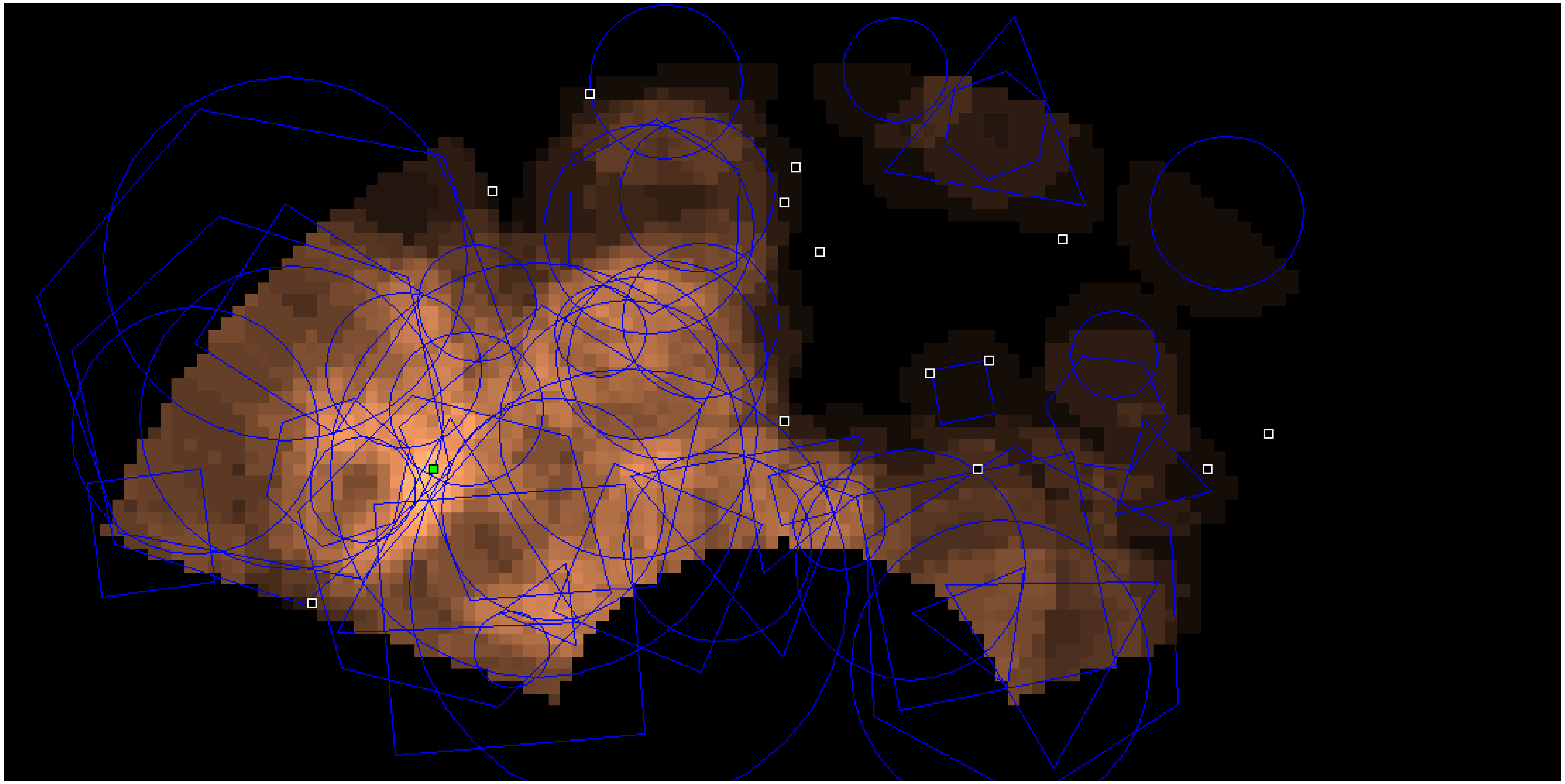
Initial Stage

BLUE: Inference Engine generates samples from space of polygons / circles
COPPER: Inquiry Engine computes entropy map of predicted measurement results



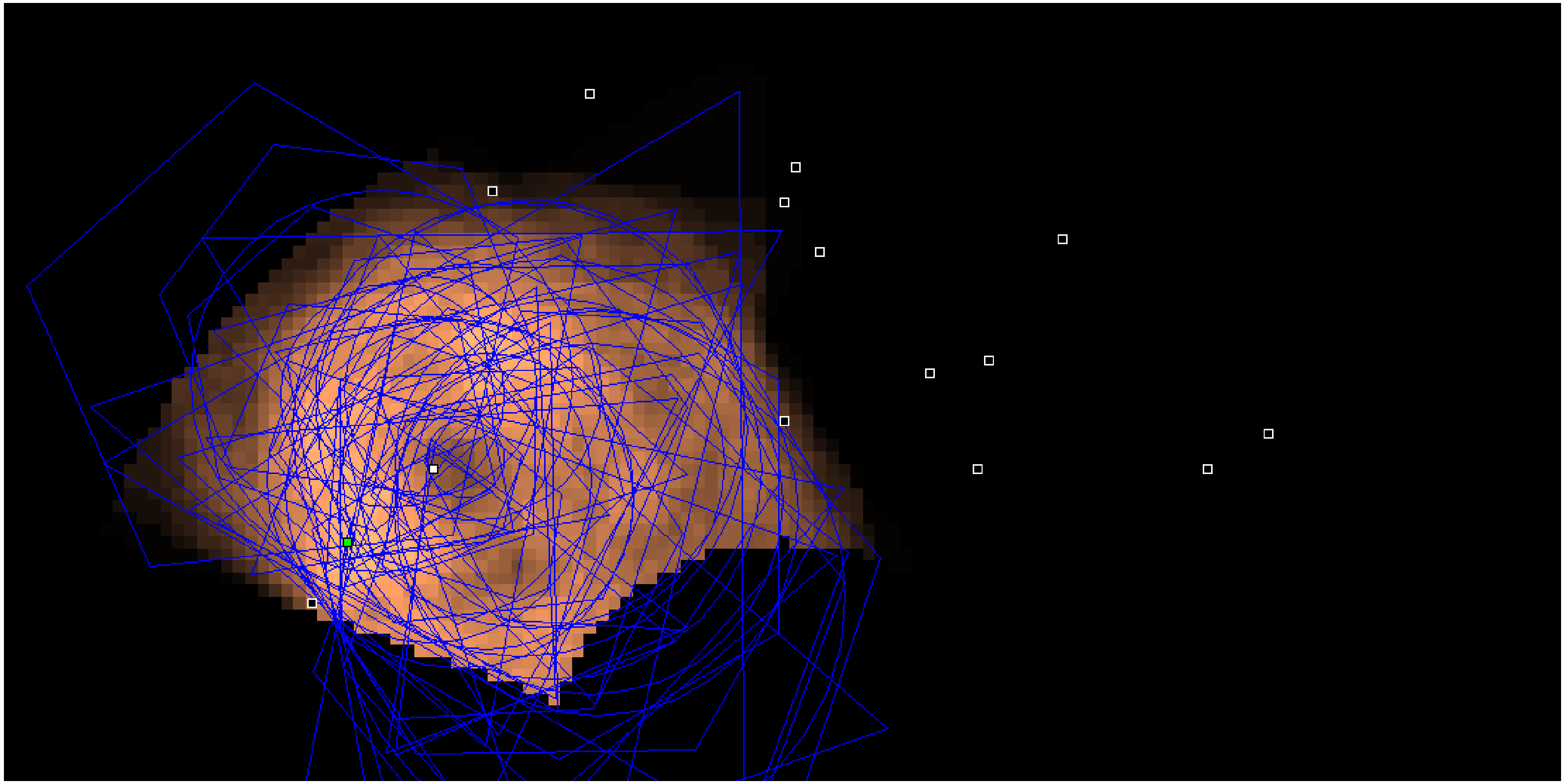
With little data, the hypothesized shapes are extremely varied and it is good to look just about anywhere

After Several Black Measurements



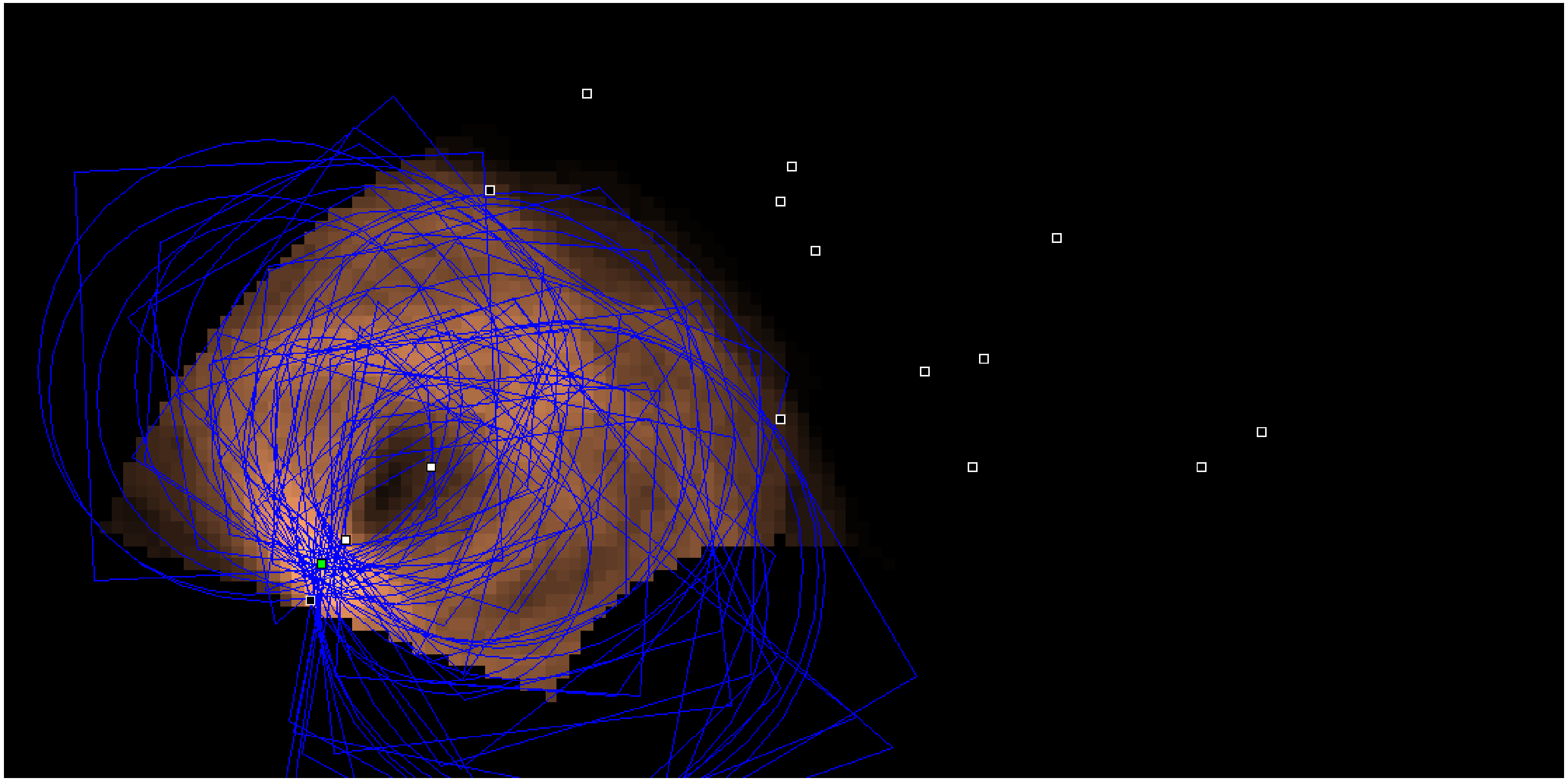
With several black measurements, the hypothesized shapes become smaller
Exploration is naturally focused on unexplored regions

After One White Measurement



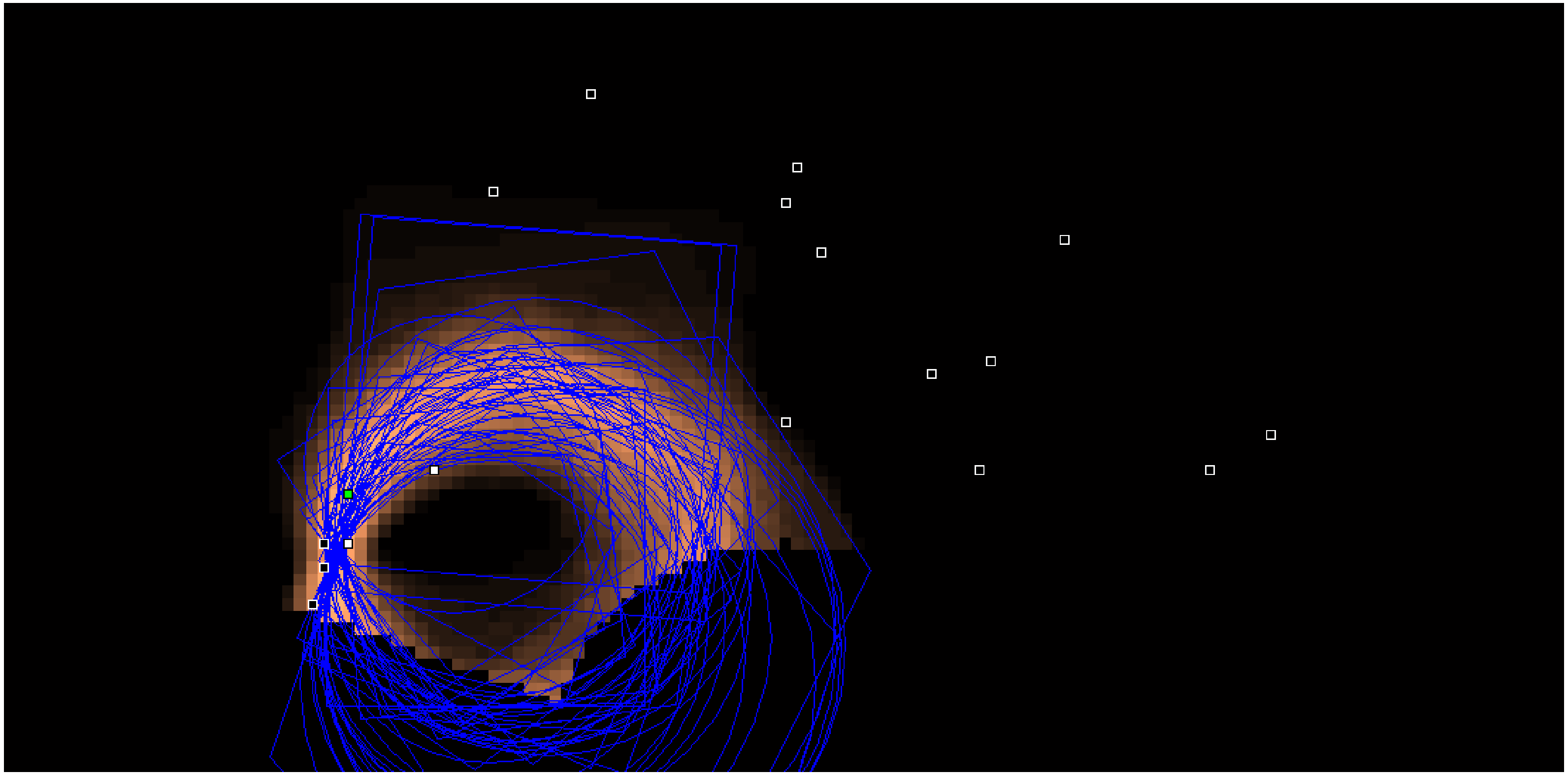
A positive result naturally focuses exploration around promising region

After Two White Measurements



A second positive result naturally focuses exploration around the edges

After Many Measurements



Edge exploration becomes more pronounced as data accumulates.
This is all handled naturally by the entropy!

Current Research

Generalize the Inference and Inquiry Engine technology to a wide array of scientific and robotic applications.

- Complex Urban Mapping
- Modeling Ephemeral Features
- Sensor Web Deployment with Swarms
- Autonomous Instrument Placement
- Autonomous Experimental Design

'Am I already in the shadow of the Coming Race? and will the creatures who are to transcend and finally supersede us be steely organisms, giving out the effluvia of the laboratory, and performing with infallible exactness more than everything that we have performed with a slovenly approximativeness and self-defeating inaccuracy?'

George Eliot (Mary Anne Evans),

The Impressions of Theophrastus Such, 1879.

Special Thanks to:

John Skilling
Ariel Caticha
Janos Aczél
Keith Earle
Philip Erner
Deniz Gencaga
Philip Goyal
Steve Gull
Jeffrey Jewell
Carlos Rodriguez

Funded in part by:

University at Albany Faculty Research Development Award

NASA AIST-QRS-07-0001
NASA 05-AISR05-0143