



deap[®]

2015

The Theory of Big Information

(AN INTRODUCTORY GUIDE)

THOMAS HAZEL

INTRODUCTION



With the advent of the Internet, cloud, mobile, and all things connected with IoT, technology is generating an epic amount of information. Over the last 15 years, the **gravity** of this information has reached a **mass** that is clearly igniting the Big Bang of Big Data. IDC has stated that 90% of the world's data has been created in the last two years with 50 billion connected "things" predicted by 2020¹. To deal with this information "Big Bang", data theory needs a new perspective.

¹ <http://share.cisco.com/internet-of-things.html>

THE BIG BANG OF INFORMATION

From Big Bang to Big Data, the physical universe and information systems (as a whole) hint to their expansion. This expansion from a cosmological perspective is based on Inflation, a well-accepted hypothesis. However, there does not seem to be an equivalent term or study for Big Data- though basic observation and intuition would say technology begets more information and more information begets new technology. Proportionally speaking, Moore's Law set forth that as size and cost of technology decreases so scale and adoption increases, thus Q.E.D. information will inevitably expand.

At the outset, this expansion of information has tremendous potential, deriving insight and value. However, it also has monstrous ramifications if not tamed. This guide will outline a new theory on how to consume and digest this potential. For purposes of this journey, let's define this expansion by taking the basic definition of "**Inflation**" and adding the word "**Information**," to make a new Big Data term "**Information Inflation**," meaning:

The logical expansion of information by ways of connecting information to produce knowledge that yields new information and/or technology (via self-similar growth)

The reasoning for a Big Bang analogy basically comes down to similarities, similarities that allow for Big Bang concepts to be used interchangeably, or at least as a point of reference, to construct a meaningful Big Data (i.e., Big Information) theory. This guide will outline this theory and how it relates to Big Data demands. First, let's get down to fundamentals before deriving new concepts.

²http://en.wikipedia.org/wiki/Moore's_Law

THE AFFINITY OF PHYSICAL AND LOGICAL UNIVERSES

Before going deep into this new Information Inflation concept, one might need a deeper understanding of this analogy. Like the Hadron Collider's mission at CERN (to find the smallest representation of matter and its relationships in the physical universe), this guide's mission is similar but with one distinction: It models a logical universe. Unlike the Collider, one does not need billions of dollars to find and describe the fundamentals of Big Data. However, these fundamentals continue to have similarities with the study of the Big Bang. It all centers around the science of modeling this expansion, none of which would be possible without the groundbreaking works of Claude Shannon.³

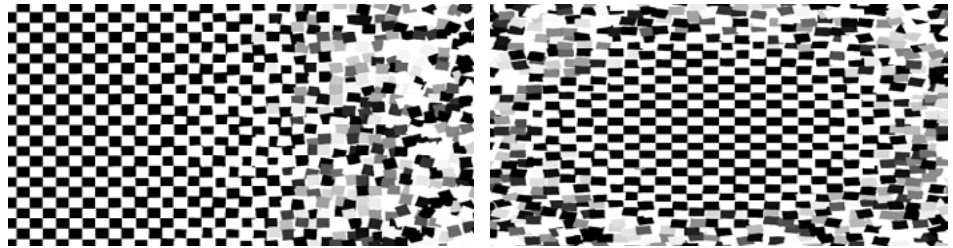
The "affinity" between the Big Bang and Big Data stems from Shannon's initial work in 1935 where he formulated that boolean algebra and binary arithmetic could construct and resolve any logical relationship (i.e., computers, algorithms), but most notably his landmark work in 1948 around Information Theory. However, the relationship between the Big Bang and Information Theory can cause some confusion with respect to the definition of Entropy and needs some clarification.

In the physical universe, entropy refers to the universal tendency for systems to become less ordered over time, a principle based on the second law of Thermodynamics. However, in the logical universe, entropy refers to how ordered (or predictable) a stream of data is. Within this guide, the latter definition will be used. However, the nature of physical entropy and some of its theoretical properties will still be overlaid.

³http://en.wikipedia.org/wiki/Claude_Shannon

THE AFFINITY OF PHYSICAL AND LOGICAL UNIVERSES

The Hadron Collider studies the physical universe by smashing subatomic particles to create the smallest Elementary Particles. Yet physicists are not satisfied with just observing these elements. They tirelessly try to split them into even smaller substructure. In the logical universe there is only one elementary particle and that's a conceptual block of **data**.



FUNDAMENTAL BUILDING BLOCKS OF THE LOGICAL UNIVERSE

AXIOM - N°1	Data (symbol) an abstract singular concept to be viewed as the lowest level of abstraction, from which information and knowledge are derived. This is not to be confused with machine-generated data that has context and/or structure.
AXIOM - N°2	Information (to inform) a single unit of contextual data and/or sequence of data described in some well-formed structure. By connecting two or more elements of Information, new information can be derived (via self-similar growth).
AXIOM - N°3	Information Entropy (disorder/uncertainty) measurement of uncertainty of random data from a source stream, characterized by the probability distribution of the samples drawn.

These definitions are fundamental building blocks and will be used to derive more complex concepts and terms throughout this journey. At this point, some self evident relationships can be deduced and it starts with the connection between data and meaning, where meaning is an inverse derivative of Information.

From a basic level, data has no meaning since it has no context or structure. Conceptually data can travel arbitrarily through the logical universe and never obtain meaning. There are two ways for data to achieve meaning: when two or more data are structured together or when the source (i.e., cause) of data is known. A simple way to illustrate these principles is to represent data as a single bit of storage in a computer's disk drive. From a naive observer, the drive is a random distribution of bits having no relationship to one another with no apparent organization. Yet in a working computer, this entropy is an illusion since the computer knows how to interrupt these bits as logical structure/ instructions and thus provides meaning to this data.

Even one bit has meaning if the source of the data is known. Imagine that the computer writes to the same location on the drive to indicate that the computer is on or off. The symbol "1" for "on" and the symbol "0" for "off". Though this data is not structured to any other data, the source of where this symbol comes from gives it meaning to the observer.

The relationship between data and information is really the transition from one perspective to another. In other words, once data has meaning (source or structure) it becomes information.

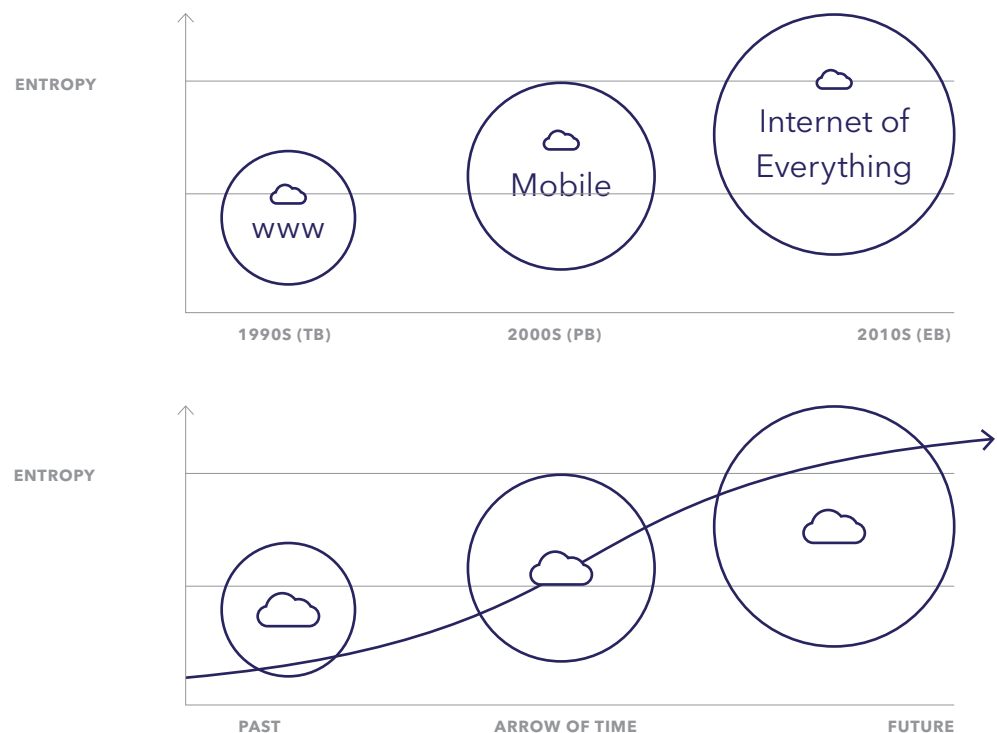
There is a theoretical possibility that adding more data to information could destroy meaning. However, this will be rejected for this journey due to the principles of Arrow of Time with respect to entropy (see: Entropy and [Arrow of Time](#) in a Logical Universe).

And finally, the relationship between information and entropy would seem down right destructive with respect to obtaining meaning. However, entropy is a measurement of disorder/uncertainty and not the loss of meaning. In other words, [Entropy \(Information Theory\)](#) always relates to some statistical model working to predict some future. Therefore, entropy can be seen as an important metric to be used to derive meaning from past, in present, for future data and ultimately order/certainty (e.g. codebreaking).

ENTROPY AND ARROW OF TIME IN A LOGICAL UNIVERSE

As mentioned earlier, knowing the source of data provides meaning. If this source continues to produce data, more meaning is derived via entropy modeling. However if a source changes so can its meaning. If the source is unknown, such data will add confusion and thus destroy meaning. To illustrate this point, let's go back to the disk drive analogy where a computer writes a "1" or "0" depending on its state of "on" or "off," respectively. If some arbitrary (i.e., unknown) source writes to the same location, the meaning of the original source would be lost. In other words, when the computer writes "1" to indicate the "on" state and the unknown source then writes "0", the observer would believe the computer to be "off."

As a source produces data, a sequence is formed and can be seen as a stream flowing in a direction from newest to oldest. This flow can be seen as a list of events chronicling the passage of time. Therefore, to interpret the expansion of information in a logical universe one must understand direction, thus without arrow of time, data is just a scattered distribution.



In the physical universe entropy and expansion go hand in hand, where randomness always increases due to inflation. However, if taming (stopping or reversing) physical entropy is an objective, there probably needs to be a change in laws of thermodynamics with respects to Entropy (Arrow of Time). In other words, reversing time is akin to re-assembling a shattered glass of wine.

Time in the logical universe, like in the physical, is irreversible. However, unlike the physical universe, the logical universe is a representation of abstraction. In other words, the logical universe is a closed "information system" that models data over time, like a video camera recording the shattering of a glass. The logical universe stores data describing physical reality or even another closed information system. Therefore, unlike the physical universe, a logical universe can stop, reverse and replay the stream of data giving the illusion of time manipulation.

As previously stated, the logical universe cannot stop time and so if a source continually produces data, inflation at some point will hit the physical limits of a closed information system.

ENTROPY AND RELATIVITY IN A LOGICAL UNIVERSE

The central concept of Information Theory is the measurement of entropy within a sequence of data. The pertinent aspects of measurement is in respect to pattern recognition and data compression. For instance, a stream of data can be sampled to determine the symbolic distribution across the entire (or relative portion) of a sequence. This sampling allows for patterns to be recognized and portions to be grouped (e.g. reordered) and/or compressed.








Knowing the nature of a source up front can be critical to understanding entropy of a data stream. For instance, a source that can only produce one symbol has an entropy of zero. Pattern recognition and/or compression in a stream with no entropy is still possible, but not terribly beneficial since logically it can be represented as a simple length of the sequence.

As entropy of a data stream increases so does the benefits of pattern recognition and compression. Compression typically works best on a finite amount of information. Generally speaking, the larger the amount of information compressed, the greater the reduction. However, a stream by definition is unlimited and thus not very conducive to compression. This is where pattern recognition comes into play. As patterns are identified, sequence portions can be grouped as finite amount of information and then compressed.

At exponential scale, pattern recognition becomes essential in taming information inflation. Modeling the Relative Entropy of a data stream can make or break a system's ability to handle such expansion. Imagine an information system ingesting a continuous video stream. Typically video is compressed by finite amount of information grouped by time. However, there are natural patterns that cross time boundaries. By using relative entropy, grouping can be performed in limitless dimensions.

DEEP
INFORMATION
SCIENCES
PRINCIPLES OF
INFORMATION

Now that the fundamental building blocks and their nature within a logical universe have been described , the Deep Information Sciences “Information Theorems” will be outlined.

THEOREM - N°1	Information is a sequence of Information (self-similar) that is segmented by consistent (well-formed) order.	
THEOREM - N°2	Segmented Information is addressable by the First where a sequence of Firsts is a Segment of Summarization.	
THEOREM - N°3	Summarization is an aggregate of statistical and probabilistic distribution of segments.	
THEOREM - N°4	Information of a Segment is equal distance to the sequence of Information between non-associated Summarizations.	
THEOREM - N°5	Sequenced Information is in direct relation to former and later Information of which Patterns can be Matched.	

With the fundamental axioms and information theorems characterized, the definition of the “Theory of Big Information” can be stated as such:

The “general purpose” reduction of inflation is optimally achieved through directional probabilistic distribution modeling resulting in sequential summarization of information.

“ ”

With the fundamental axioms and information theorems characterized, the definition of the “Theory of Big Information” can be stated as such: Based on this Theory of Big Information and principles around Algorithmic Information Theory, the taming of the Big Bang of Information has begun.



Deep Information Sciences provides the first plug-and-play solution that leverages machine learning to evolve MySQL databases and infrastructure for the new economy's big data needs. Deep provides the only solution that allows databases to adapt to any situation, and accelerate performance and scale while streamlining infrastructure. Deep enables companies to focus on their innovation, instead of their infrastructure, bringing compelling solutions to market at significant cost savings.

Through the power of Deep's new computer science, CASSI, companies can easily unlock the full potential of their existing infrastructures, wielding it to their competitive advantage.

Deep is headquartered in Boston's Innovation District. Learn more at www.deepis.com.

Deep Information Science

www.deepis.com
1.800.270.3580
sales@deepis.com

1 Marina Park Drive
Suite 315
Boston, MA 02210