

The background of the entire slide is an abstract, textured pattern. It consists of numerous irregular, rounded shapes in shades of purple, lavender, and gold, scattered across a light gray background. These shapes have a crystalline or organic appearance, with some showing internal detail and others being more solid in color.

エントロピーと情報理論

-- 量子情報理論入門 --

丸山不二夫

How many atoms are in a human body?

Anyone have a guess? About 10^{28} .

How many stars are in the entire observable universe?

Something like 10^{23} .

How many subatomic particles are in the observable universe? No one knows for sure — but 10^{90} is a reasonable estimate.

Scott Aaronson

*Why don't you call it entropy? In the first place,
no one understands entropy very well, so in any
discussion you will be in a position of advantage*

John von Neumann

But in his final attack Einstein pointed to something so deep, so counterintuitive, so troubling, and yet so exciting, that at the beginning of the twenty-first century it has returned to fascinate theoretical physicists. Bohr's only answer to Einstein's last great discovery—the discovery of entanglement—was to ignore it.

Leonard Susskind

はじめに

- 今回の講演は、時代とともに相貌を変える「エントロピー」という概念を、身近なものとして理解してもらうことを、一つの目的としている。
 - 19世紀に、熱力学の中で定式化され統計力学を成立させた「エントロピー」は、20世紀、情報理論の中で、「情報量」として再定義され、21世紀、物理学の基礎理論の中で「エンタングルメントのエントロピー」として、再々定義される。
 - 講演では、こうした大きな流れを、いくつかのトピックスやエピソードを中心に紹介したいと思う。限られた講演時間で、フォーカスを明確にするため、「複雑さについて」と「ディープ・ラーニングとエントロピー」の部分は、Appendixに回した。
-

はじめに

- 講演のタイトルも、ミスリーディングだったかもしれない。情報の問題に関わるIT技術者にとって、今、必要なことは、何もシャノンの「情報理論」を、改めて学ぶことではないからだ。
 - 重要なことは、「エントロピー＝情報」をキーコンセプトとして、現在、科学の世界で急速に進行中の変化を知ることだと思う。なぜなら、この変化は技術の世界での大きなイノベーションを引き起こす、実践的に大きな意味を持っているからだ。変化は、おそらく、急速に進行するだろう。「量子情報」の世界は、すぐそこに来ている。
 - 本講演を、「量子情報理論」の入門として、受け止めてもらえれば嬉しい。いつか、マルレクで「量子コンピュータ」の話をしたいと思う。
-

Agenda

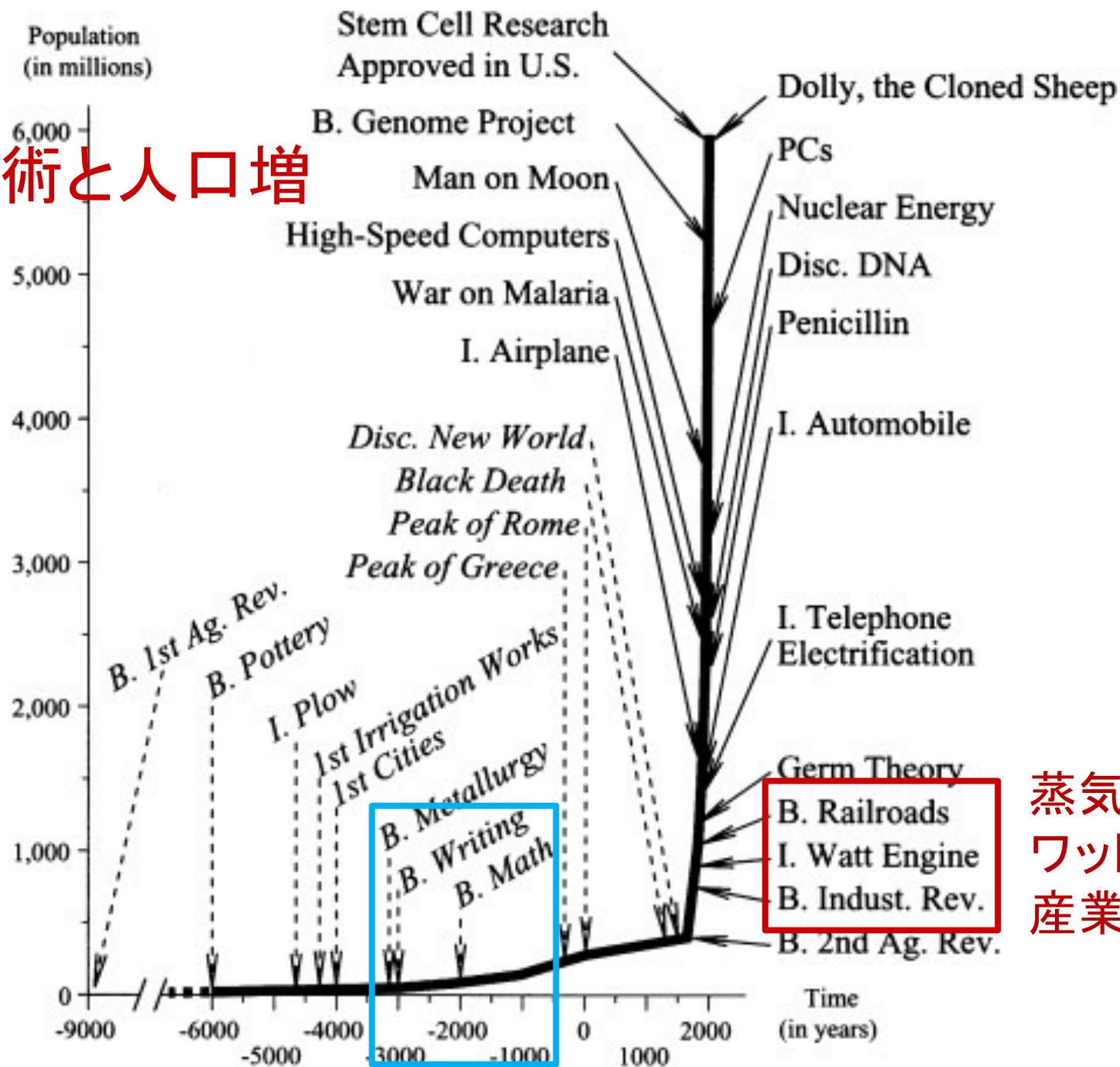
- イノベーションと科学
 - エントロピーをイメージする
 - 19世紀 ボルツマン:統計力学とエントロピー
 - 20世紀 シヤノン:情報理論とエントロピー
 - 21世紀 エンタングルメントとエントロピー
 - なぜ、今、「情報理論」なのか
 - Appendix
 - 複雑さについて
 - ディープ・ラーニングとエントロピー
-

The background of the slide is a light lavender color, decorated with numerous irregular, splatter-like shapes in shades of purple, gold, and grey. These shapes are scattered across the entire surface, creating a textured, artistic effect.

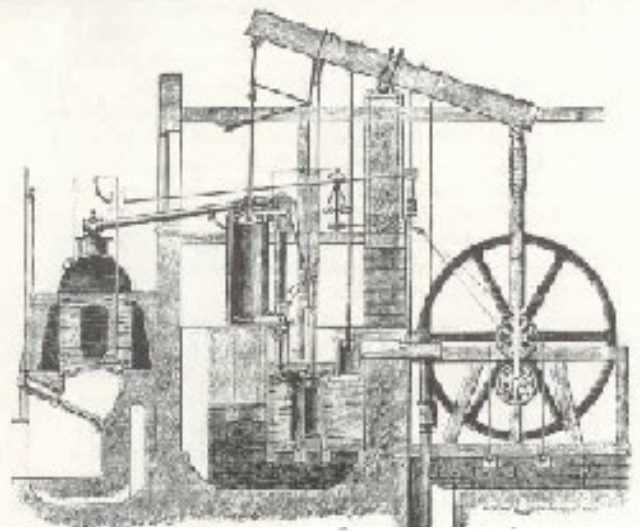
イノベーションと科学

イノベーションの歴史的進行

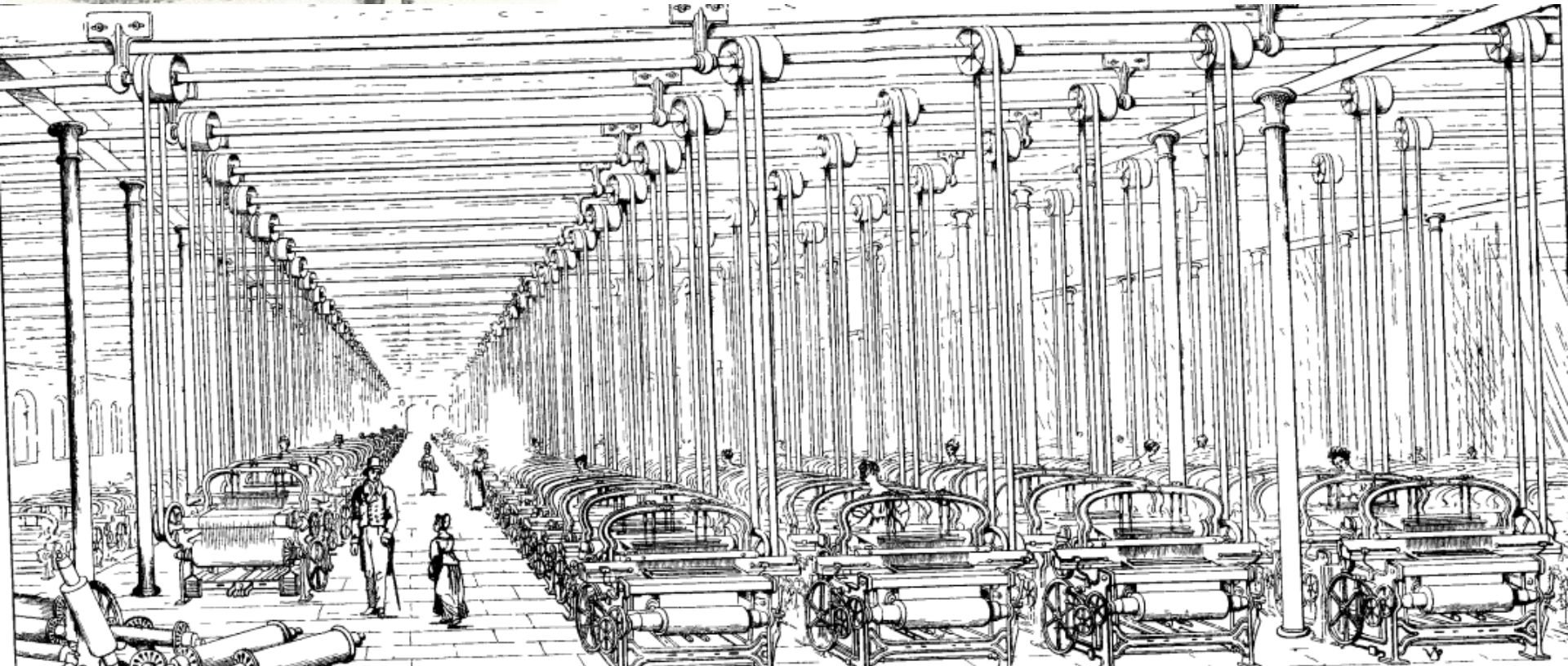
技術と人口増



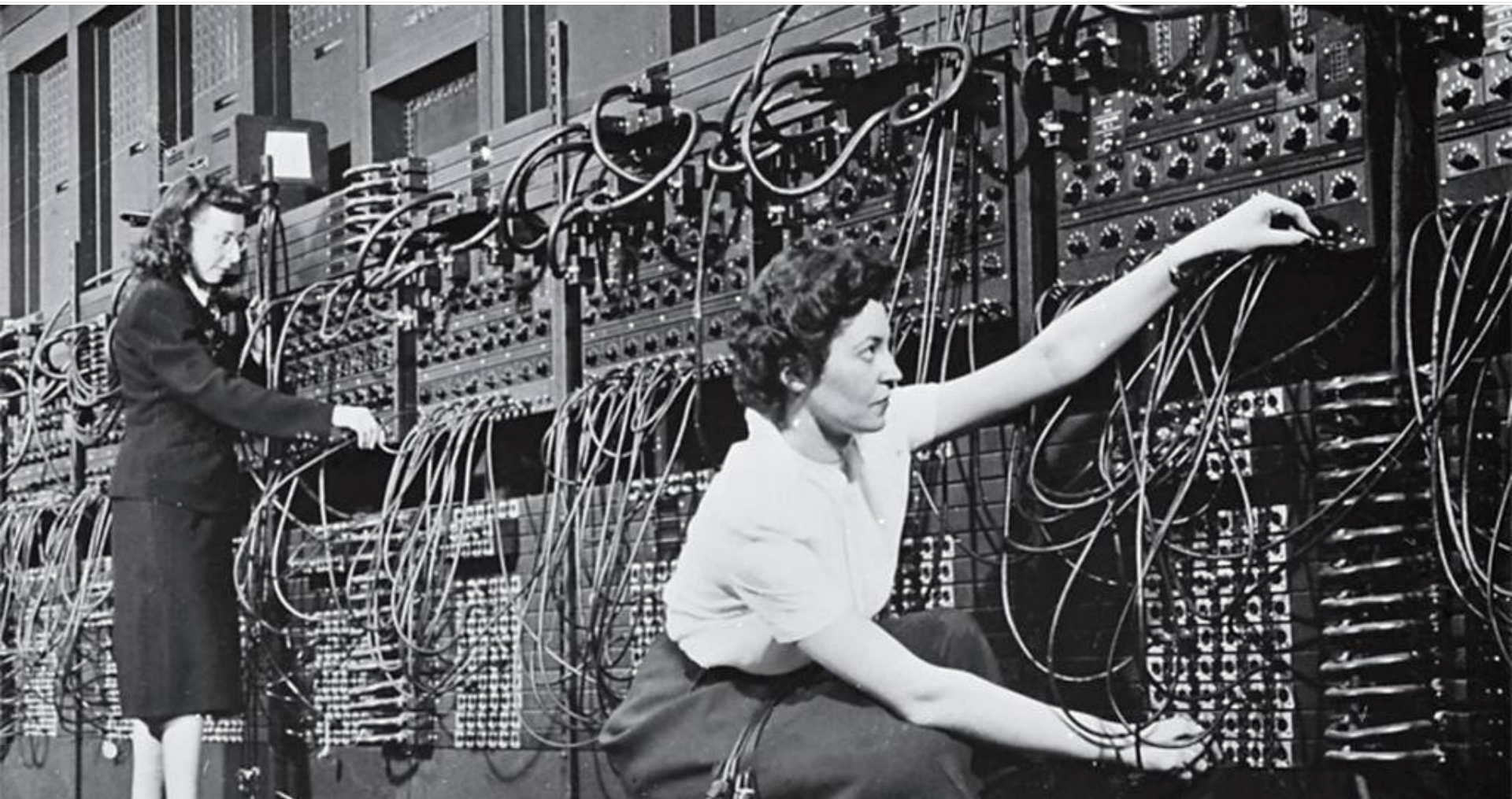
蒸気機関車
ワット蒸気機関
産業革命



18世紀 産業革命 19世紀 機械制大工場



20世紀 ENIAC



21世紀 Google

Google I/O 2013: Day 1 Technical Sessions 2



Massive Scale

21世紀 Google

Google I/O 2013: Day 1 Technical Sessions 2

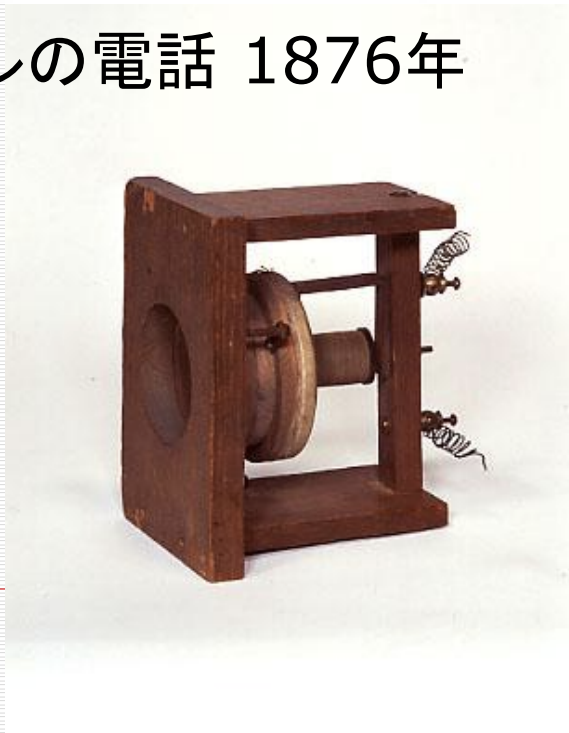


Highly Efficient Servers

ペリーが江戸幕府に送った電信機 1854年



ベルの電話 1876年



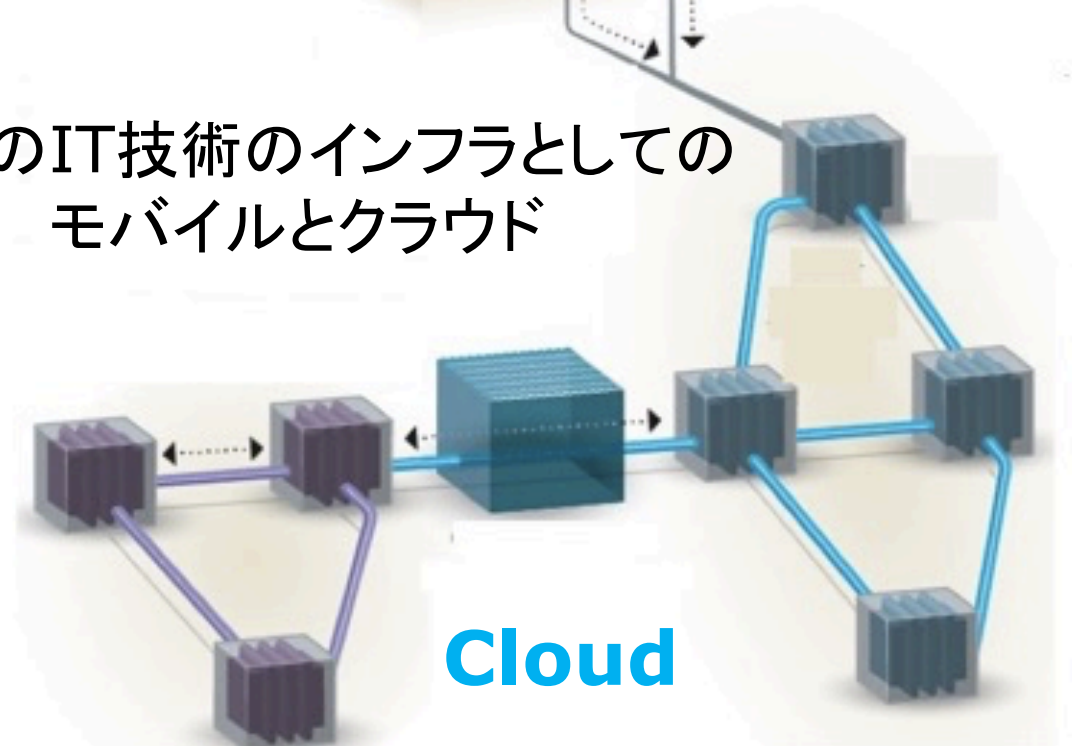
マルコーニの無線 1895年



Mobile



現代のIT技術のインフラとしての
モバイルとクラウド

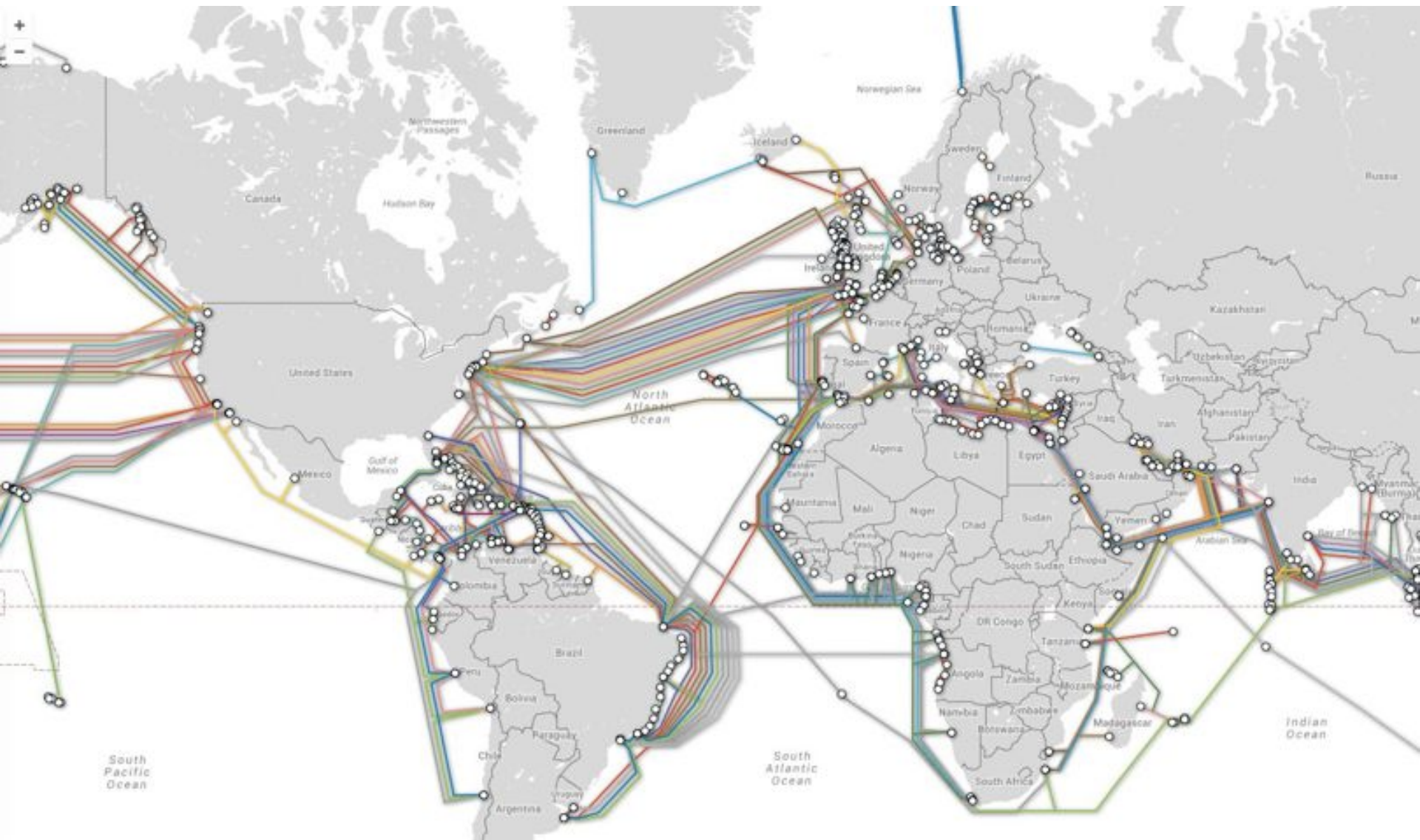


Cloud

1891年の電信ネットワーク

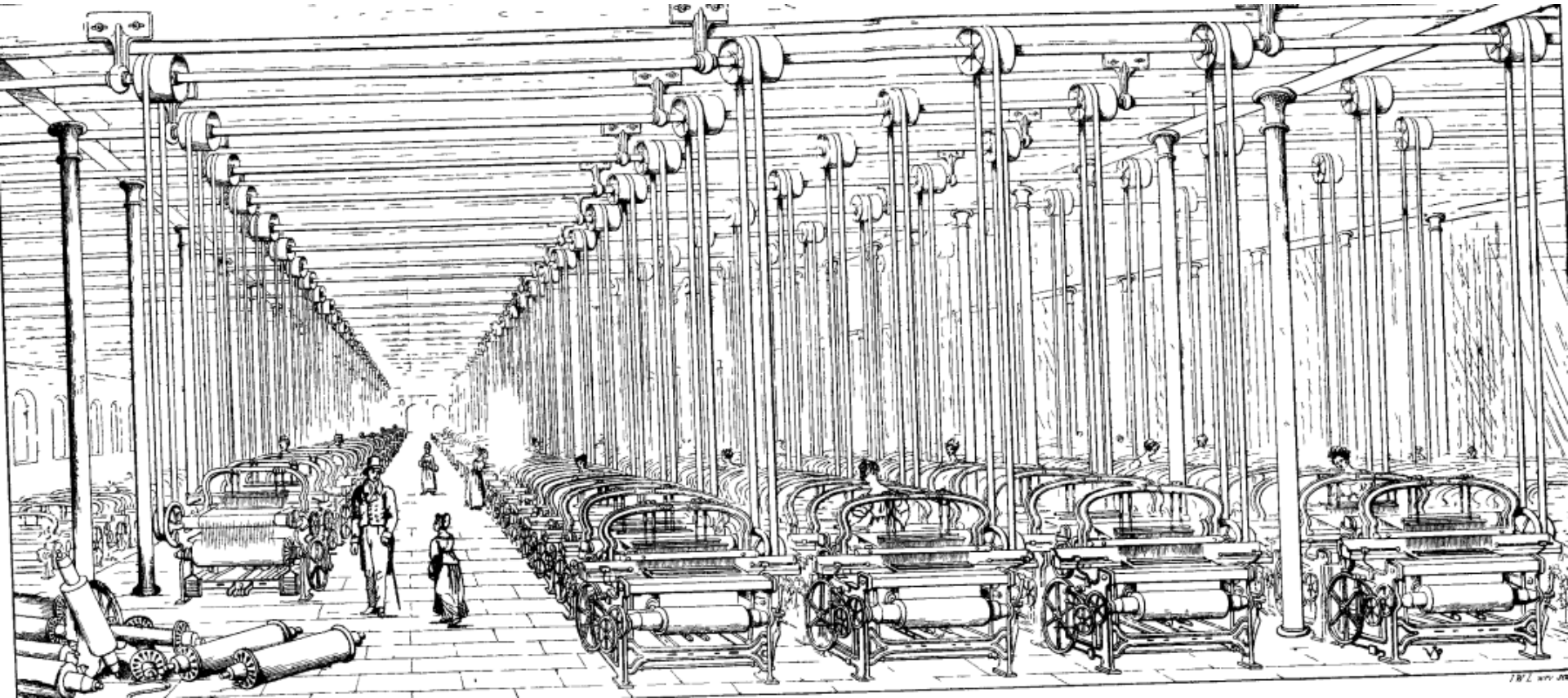


2017年のネットワーク



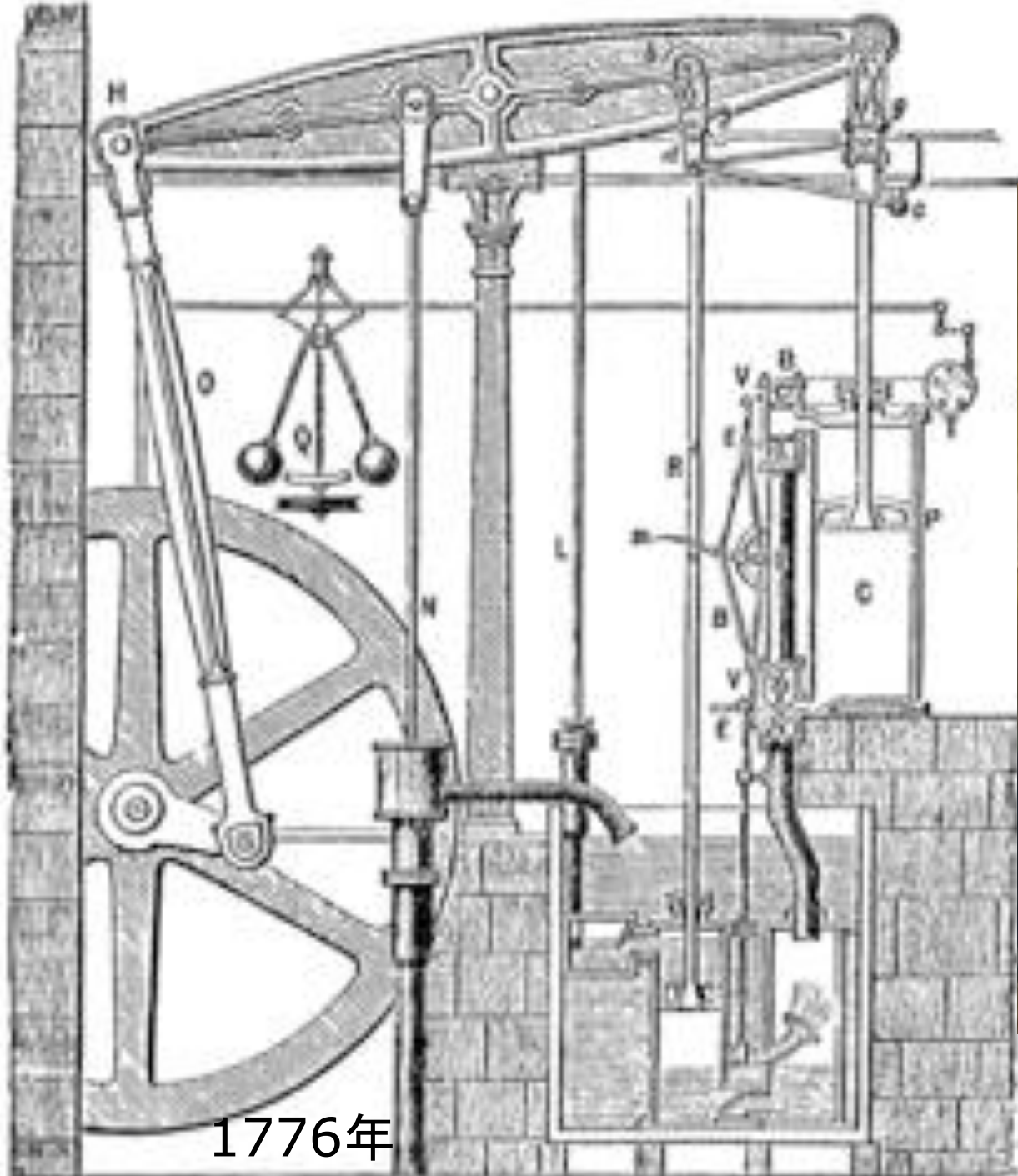
技術を基礎づける科学

蒸気機関の場合



エントロピーとは、何か？

- Entropy という言葉は、19世紀半ば(1856年?)に、Clausiusによって名付けられた。それは、「熱」や「温度」「仕事」「エネルギー」といった量に関連づけられ、「熱力学」の中で生まれた概念であった。
 - Wattの最初の蒸気機関が完成したのは、1776年。熱力学の基礎を作ったCarnotの仕事は、1828年である。彼らの実践的関心は、産業革命を可能にした当時の最先端技術である「蒸気機関」の効率化に向けられていた。
 - 彼らは、蒸気機関に熱として与えられる蒸気機関内部のエネルギーの全てが、仕事として外部に取り出せるわけではないことを、経験的に発見する。
-



1776年

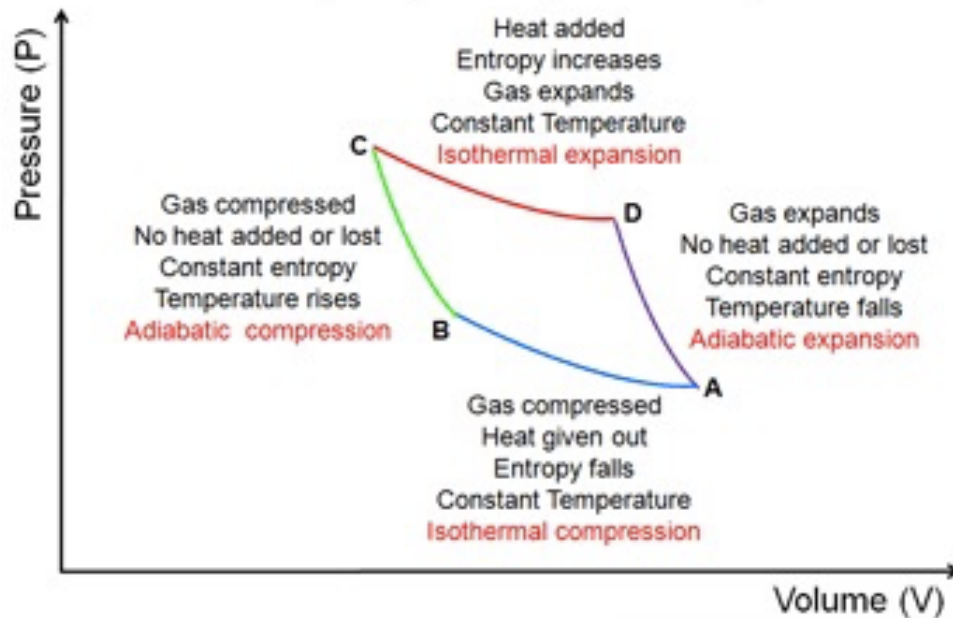


James Watt

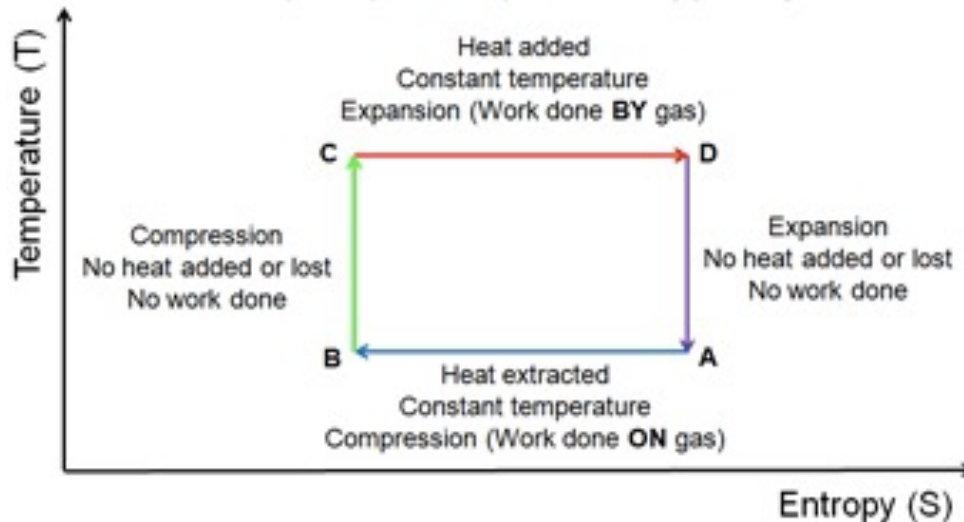
1736 - 1819

Carnot Cycle

Carnot (Ideal) Heat Cycle P-V Diagram



Carnot (Ideal) Heat Cycle Entropy Diagram



Carnot

1796-1832

エントロピー $S = \frac{Q}{T}$ 1865年

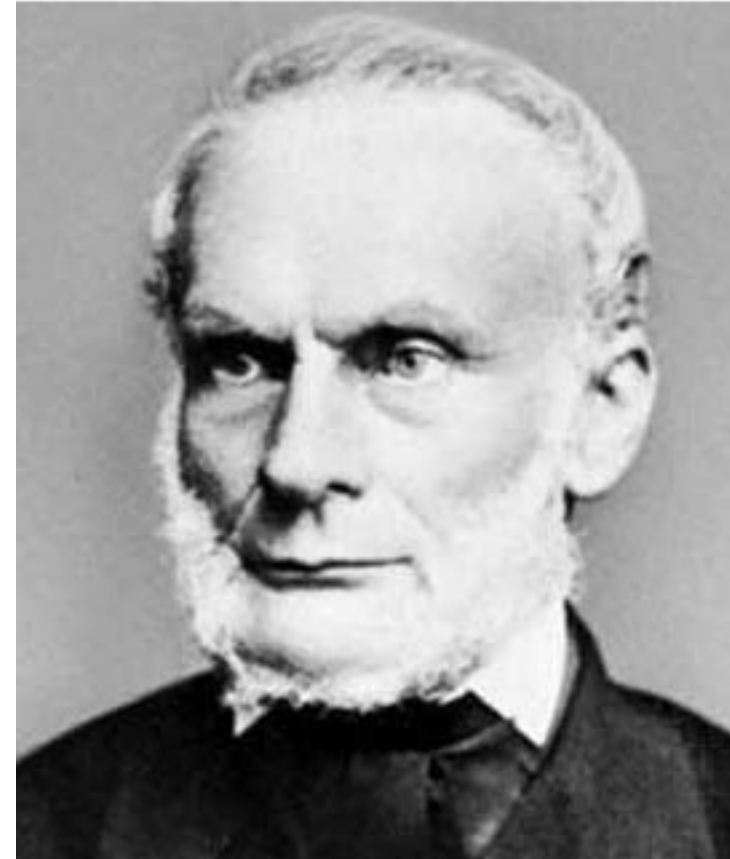
I propose to name the quantity S the entropy of the system, after the Greek word [$\tau\rho\omicron\pi\eta$ *trope*], the transformation.

I have deliberately chosen the word entropy to be as similar as possible to the word energy:

$$\sum_{i=1}^n \frac{Q_i}{T_i} \leq 0$$

$$TdS = dU + PdV$$

Rudolf Clausius

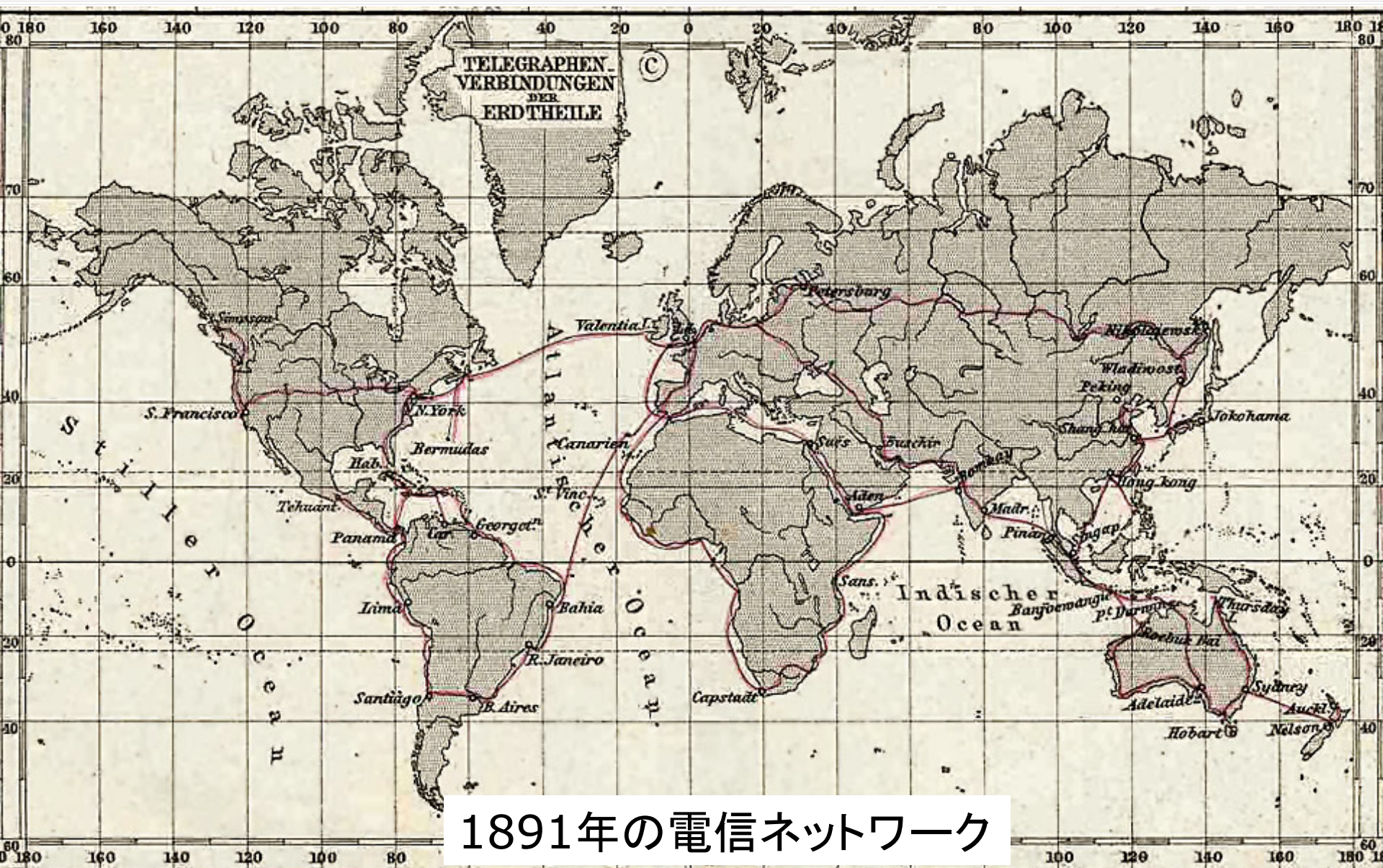


1822 - 1888

-
- 19世紀末に熱機関の熱力学を定式化した Helmholtz や Gibbs は、仕事として外部に取出せるエネルギーを「自由エネルギー」と呼ぶのだが、エントロピーは、仕事として取り出すことのできないエネルギーに関係付けられている。
 - 内部エネルギーを U 、温度を T 、エントロピー S とすると、ヘルムホルツの自由エネルギー F は、 $F=U-TS$ で定義される。
 - この式から分かるように、 TS すなわち、温度とエントロピーを掛けたものは、エネルギーである。逆に言えば、エントロピーとは、単位的には、エネルギーを温度で割ったものである。

エネルギーを温度で割ったものが、なぜ、情報の理論と関係がつくのか？ こうした関係は、古典的な熱力学の議論からは、まったく出てこないのである。 **ボルツマンが鍵を握っている。**

通信技術の場合

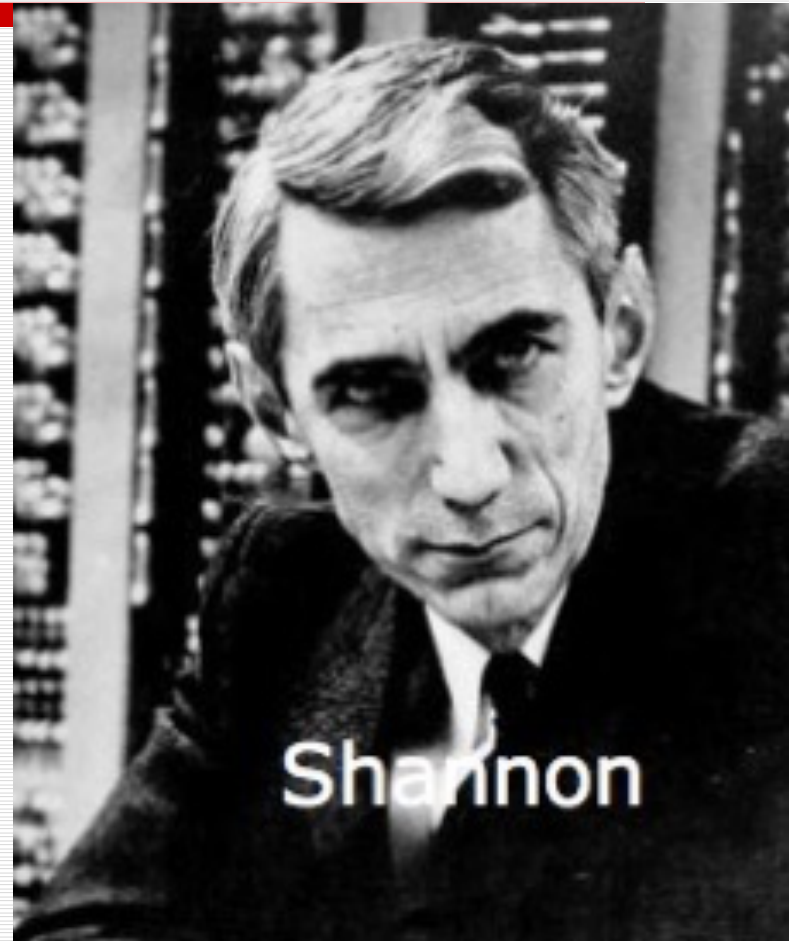


A Mathematical Theory of Communication

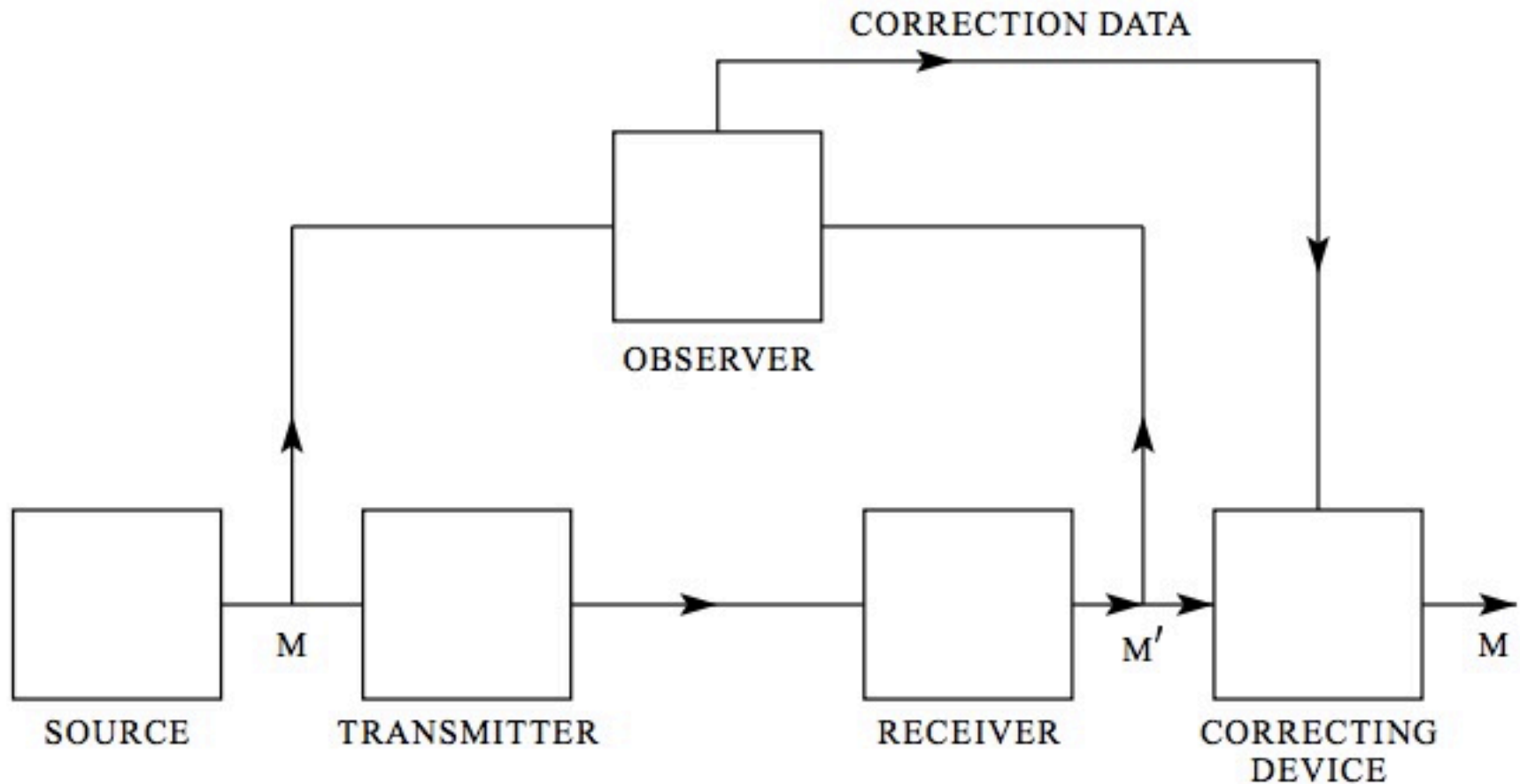
C. E. SHANNON

<http://www.qsl.net/n9zia/pdf/shannon1948.pdf>

1948年

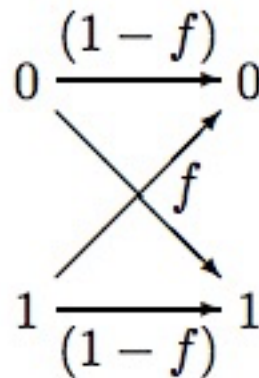


PART II: THE DISCRETE CHANNEL WITH NOISE

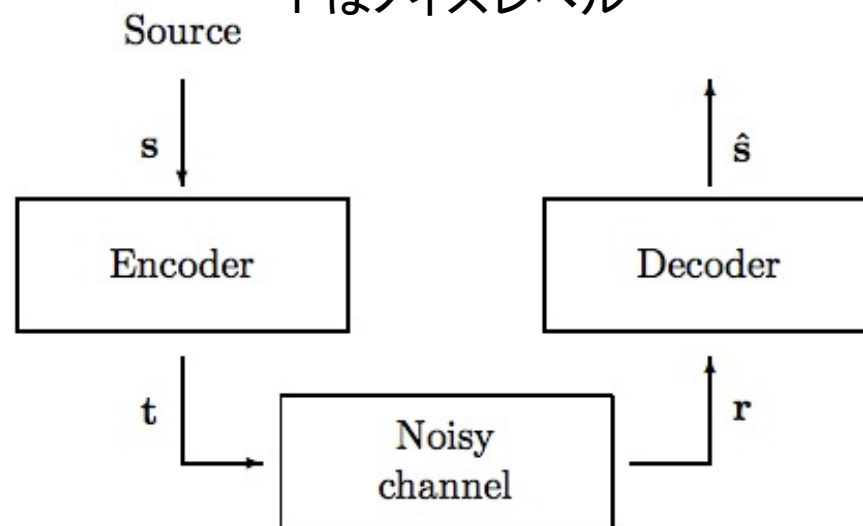


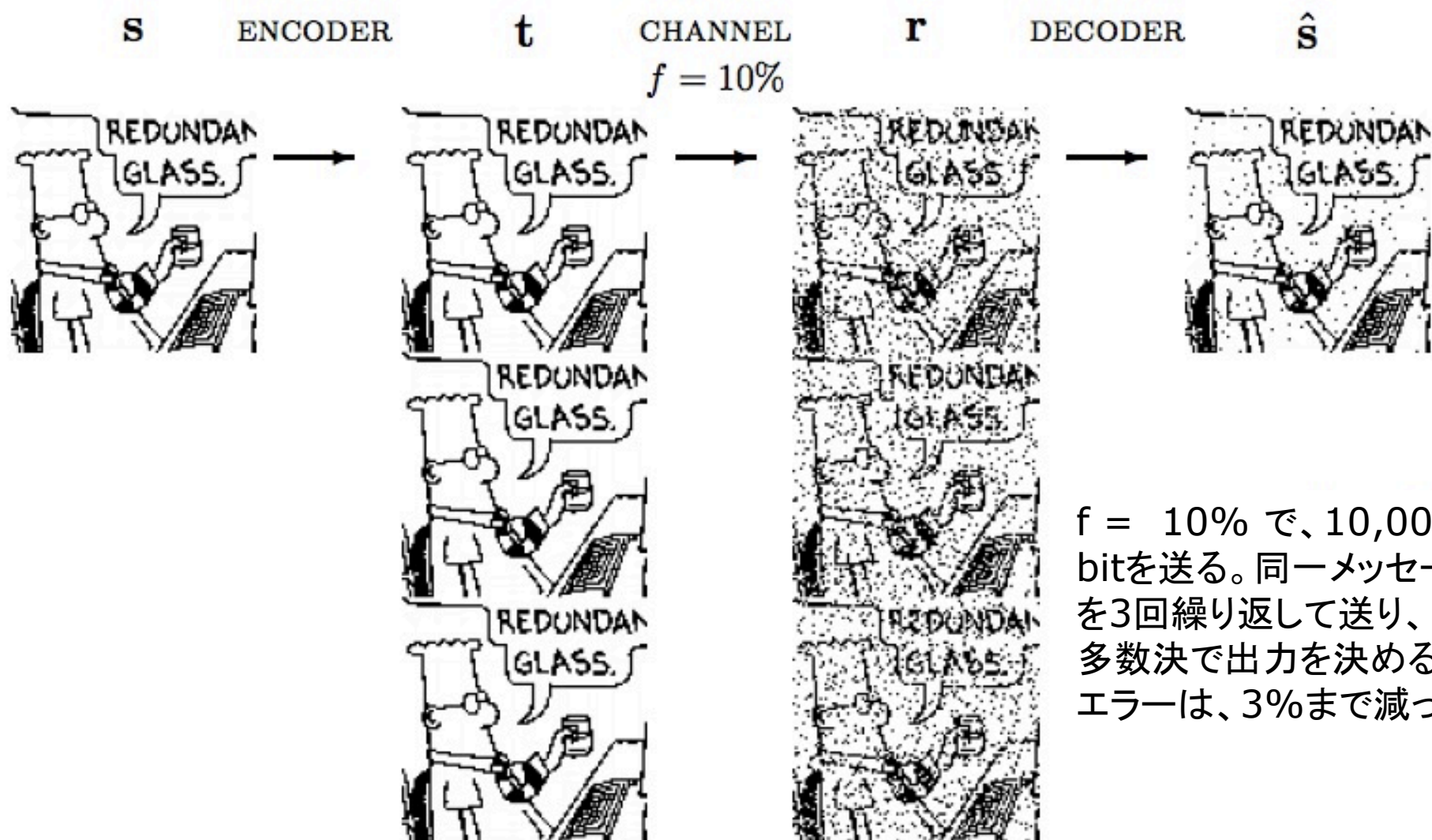
$$\begin{array}{c}
 x \begin{array}{c} 0 \xrightarrow{\quad} 0 \\ \quad \times \\ 1 \xrightarrow{\quad} 1 \end{array} y
 \end{array}
 \begin{array}{l}
 P(y=0|x=0) = 1-f; \quad P(y=0|x=1) = f; \\
 P(y=1|x=0) = f; \quad P(y=1|x=1) = 1-f.
 \end{array}$$

二値対称チャンネル

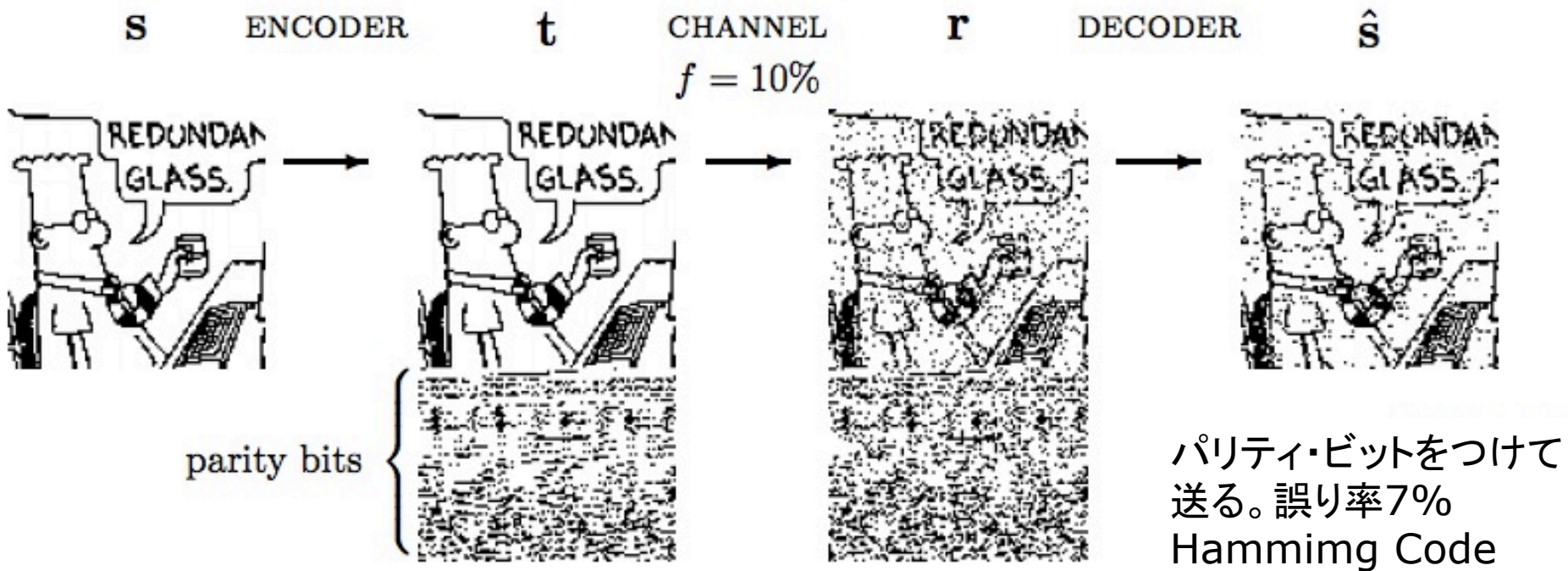


f はノイズレベル





$f = 10\%$ で、10,000 bitを送る。同一メッセージを3回繰り返して送り、多数決で出力を決める。エラーは、3%まで減った。



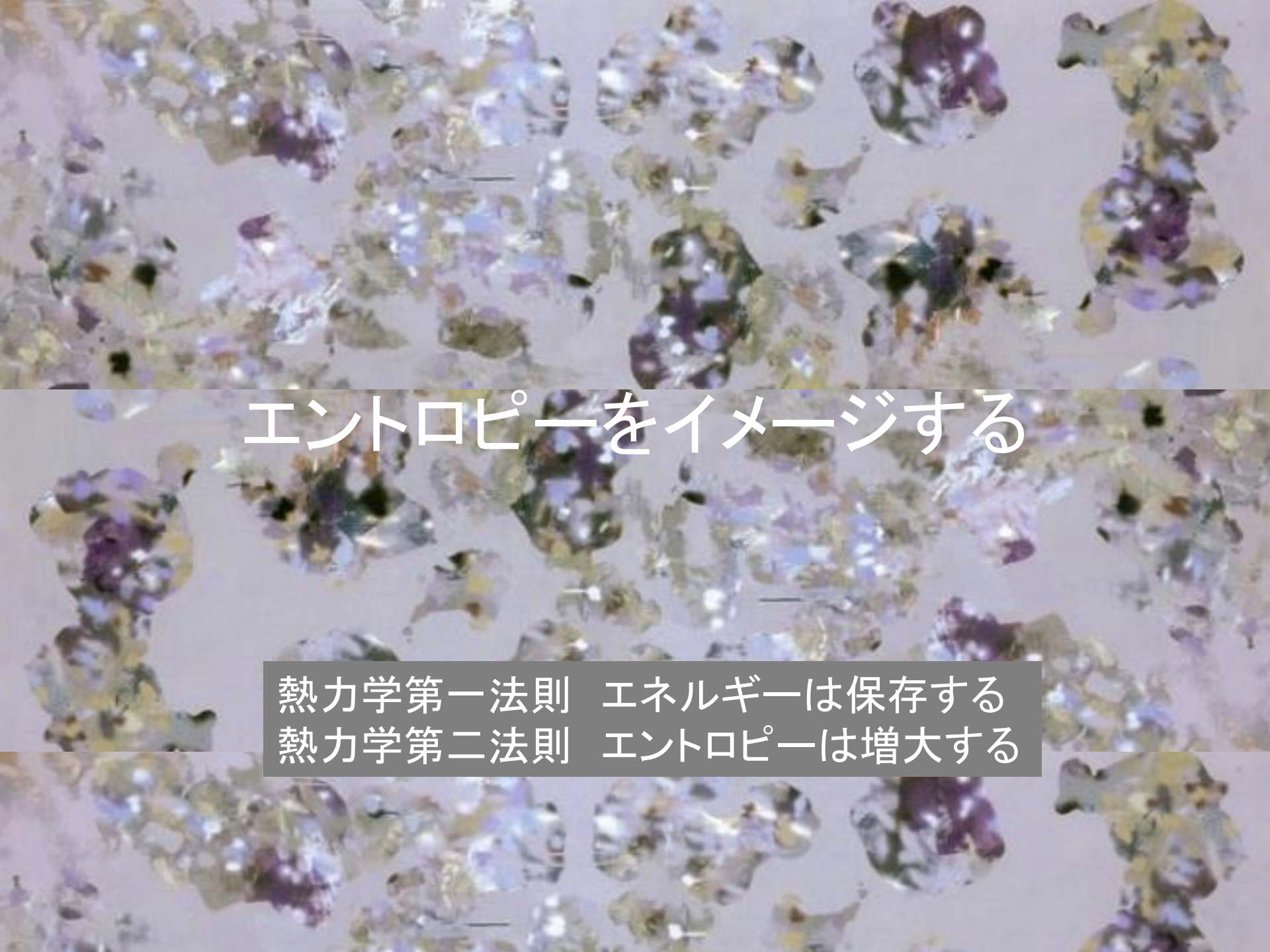
ハードディスクの読み書きに必要とされる精度
 10^{-15} 以下。

シャノンの基本的関心

通信の基本的な問題は、一方のポイントで選択されたメッセージを、他方のポイントで、正確に、あるいは、ほぼ正確に、再生することである。 (Claude Shannon, 1948)

- どうしたら、不完全でノイズのある通信チャンネルで、完全な通信を実現できるか？
- 二値対称チャンネルでのエラー訂正コード
- どのようなパフォーマンスを最良のコードは達成できるか？

なぜ、このような問題意識が、熱力学的なエントロピーと、結びつくのか？ 後の節で説明する。



エントロピーをイメージする

熱力学第一法則	エネルギーは保存する
熱力学第二法則	エントロピーは増大する

朽ち果てる？

- アマゾンの密林に新車のBMWを置き、500年もすれば、鉄の塊になる。さらに500年放置すれば、鉄サビの山ができるだろう。鉄の形は、時間とともに変わるのだが、いくら待っても、それが元の新車のBMWに戻ることはない。
 - それは、なぜか？
-

朽ちたものは元に戻らない



「人々から置き去りにされ、さび付き朽ち果てた戦車や戦闘機などの写真集」
http://gigazine.net/news/20110526_remains_of_war/

朽ちたものは元に戻らない

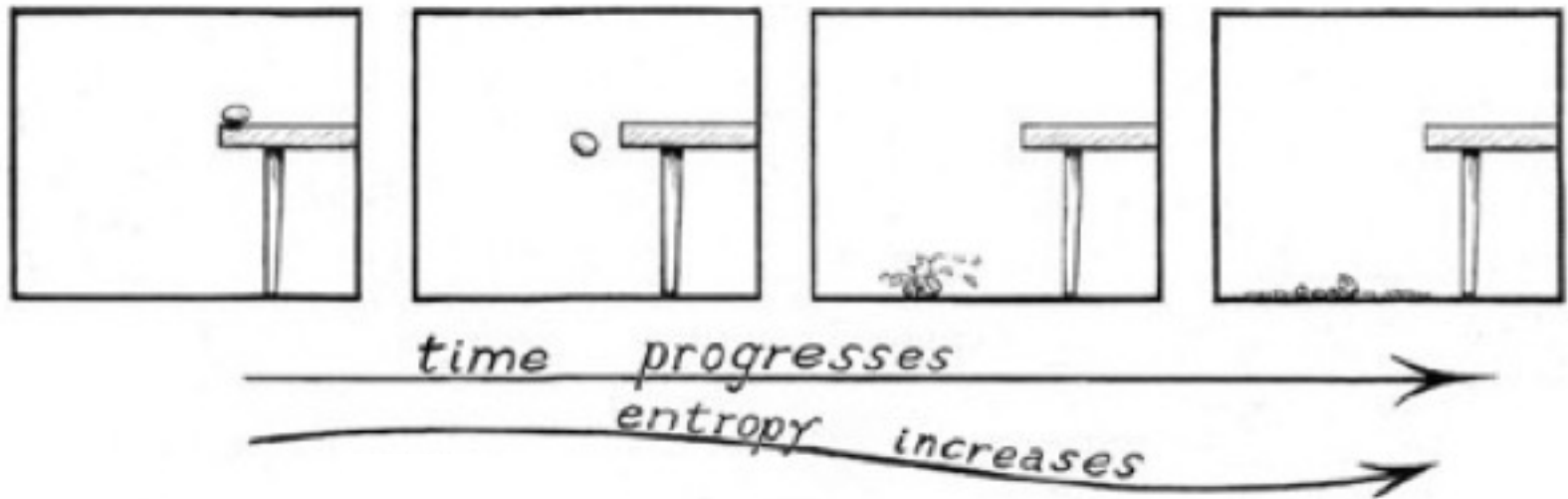


朽ちたものは元に戻らない

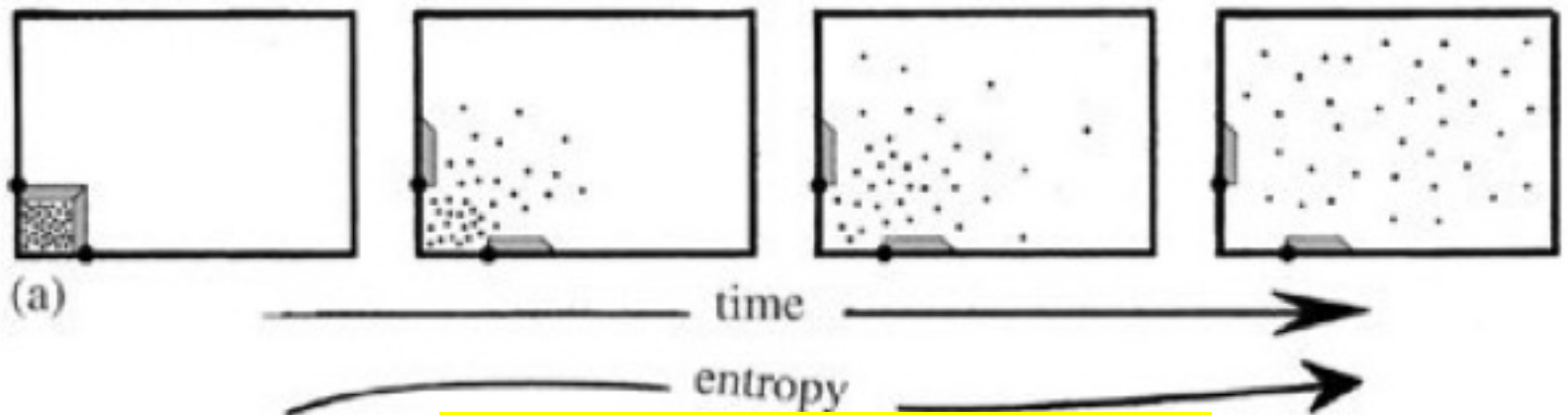


落ちた卵、拡散するガス

テーブルから落ちて割れた卵は、元には、戻らない



ガス容器から部屋に拡散したガスは、容器には戻らない



Roger Penrose "Cycle of Time"

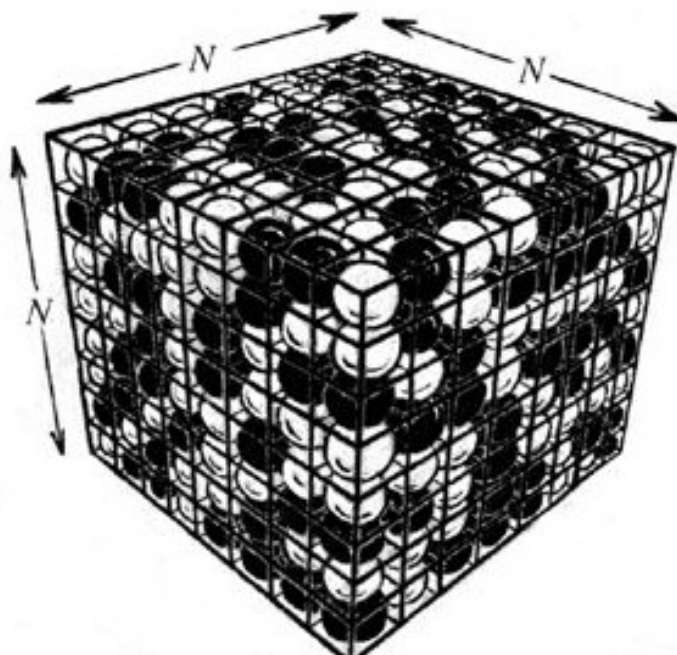
二色のペンキを混ぜる

- 同じ密度の、赤と青のペンキを、「よくかき混ぜて」みよう。そうすれば、紫の色になる。なぜ、右のようにならないのだろう？ 「よくかき混ぜた」から？



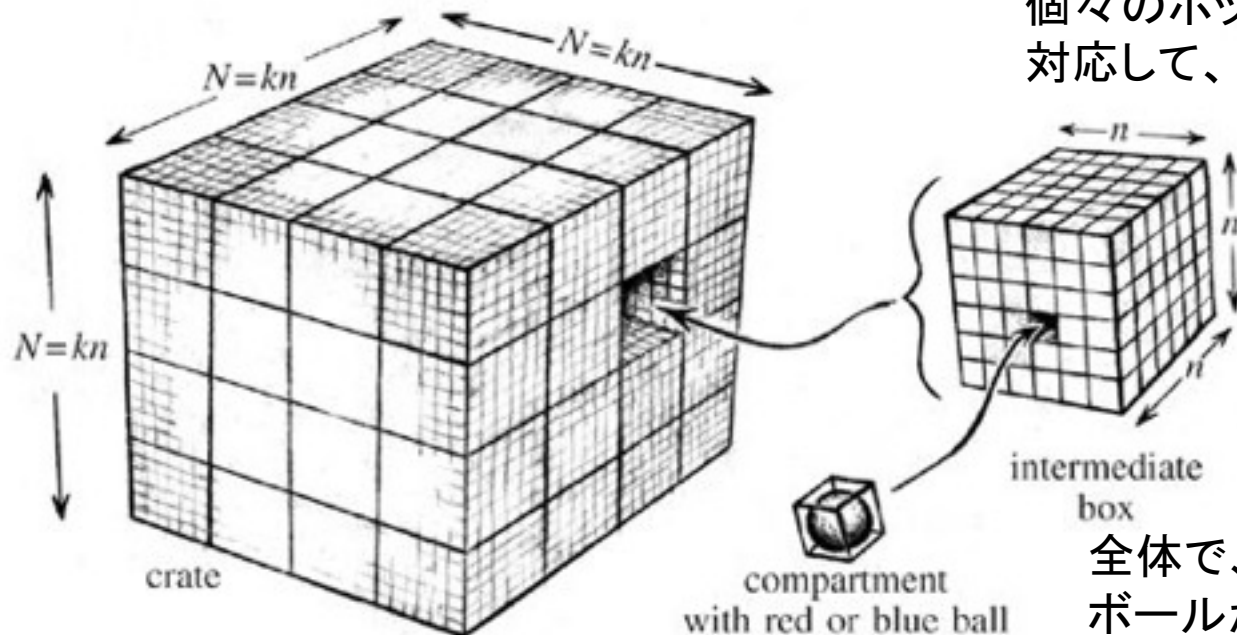
- ちなみに、青のペンキの比重が、赤のペンキより重ければ、時間がたてば、分離して、右のようになることはありうる。

赤のボール r と
青のボール b が
 $N \times N \times N$ のボックス
に入っているとする
ただし、 r と b は、同じ
数だとしよう。



N が小さいと、
赤のボールと
青のボールは、
それぞれ、はっきり
見える。

$1 \ll n \ll N$
と、大きさを増やしていくと、
個々のボックスも、 r/b に
対応して、「色」を持つ



アボガドロ数
 6×10^{23}

全体で、 10^{24} ほどの
ボールがあるでしょう
 $N=10^8$, $k=10^3$, $n=10^5$

可能な配置の数

X

$X = 10^{23} 570000 000000 000000 000000$

$Y = 10^{46} 500000 000000$

$N = 10^{24}$ として、赤のボールと
青のボールの可能な配置の数を
計算して見る

可能な配置の数

Y

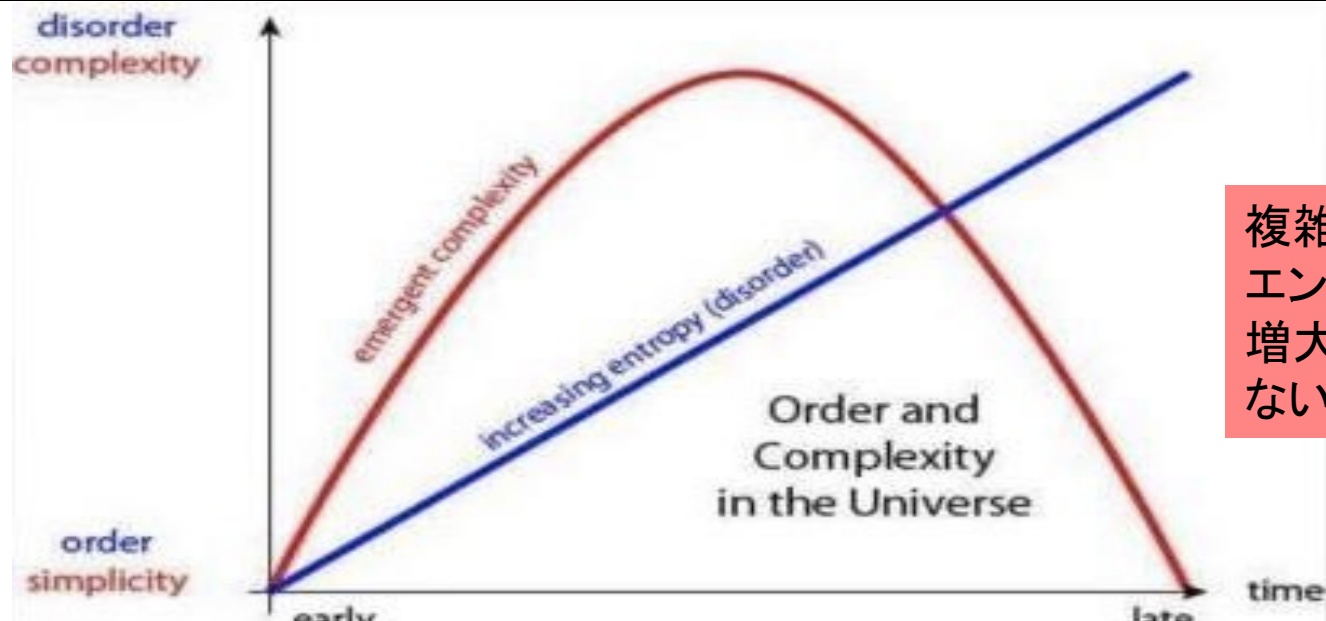
$Y = 10^{46} 500000 000000$

$Y/X = 10^{-23} 570000 000000 000000 000000$

Xである確率が、圧倒的に高い。

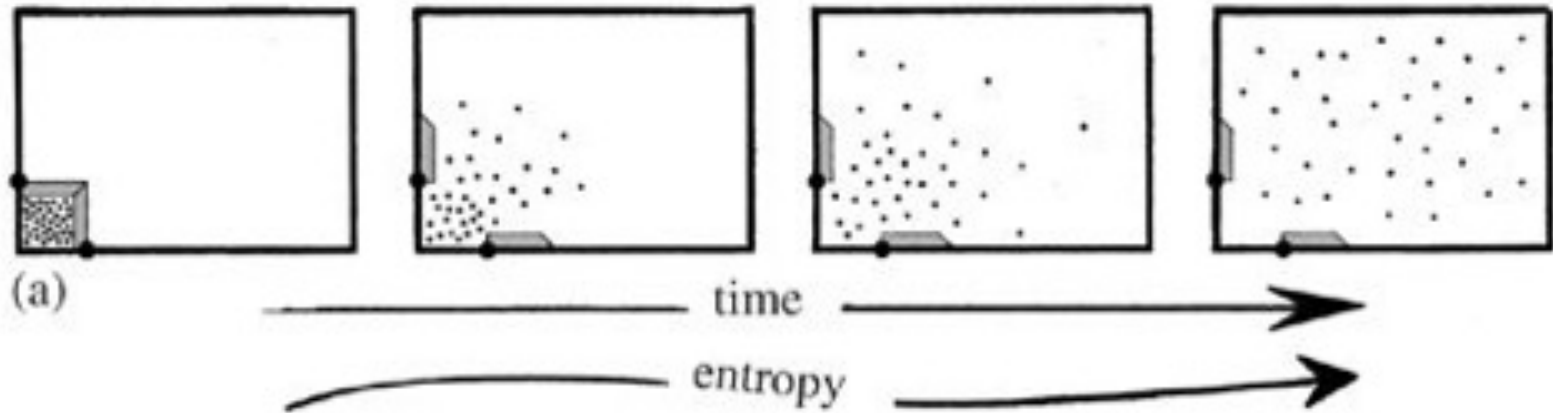


ミルクとコーヒーが混じるのを待つ

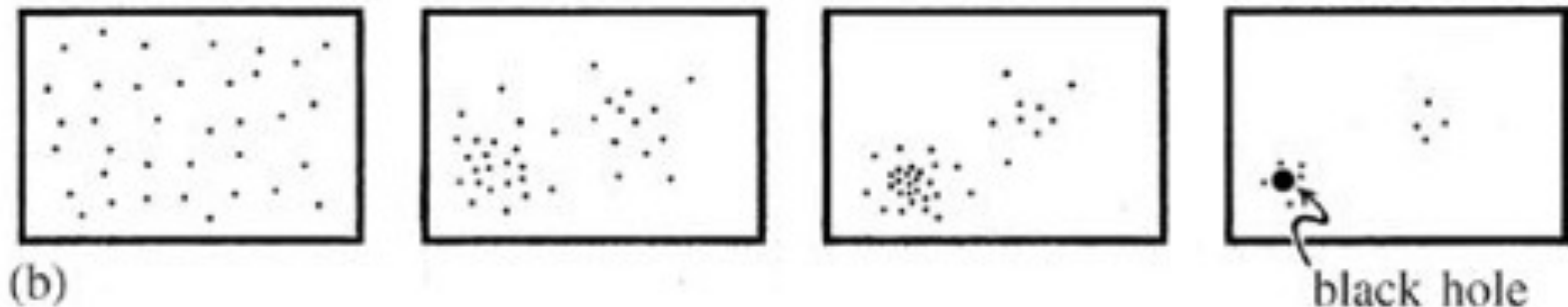


複雑さの増大と
エントロピーの
増大は同じでは
ない。

先に見た、ガスが部屋に拡散するモデル
これで、エントロピー増大のイメージを持つことができる。



ところがだ。部屋ではなく、宇宙で考えよう。
宇宙に拡散しているガスは、重力で互いに集まり、凝縮して星になる。
最後には、ブラックホールができて、そこにすべて飲み込まれる。



ブラックホールを考えると、最初のエントロピー増大のイメージは、
維持できなくなる。

Roger Penrose "Cycle of Time"

Copyrighted Material

HOW BIG IS

BIG

AND HOW SMALL IS

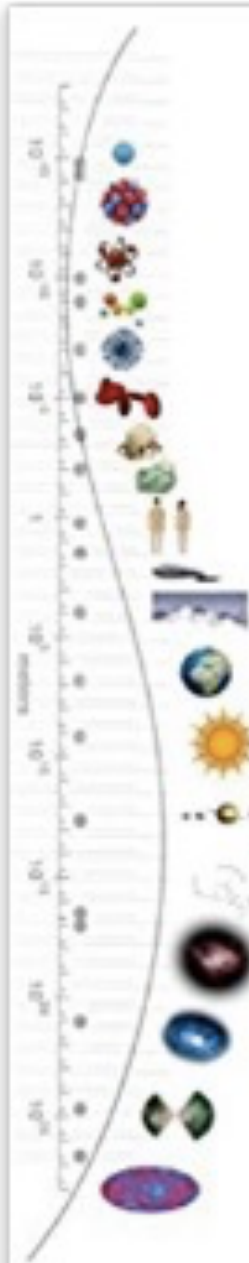
SMALL

THE SIZES OF EVERYTHING AND WHY

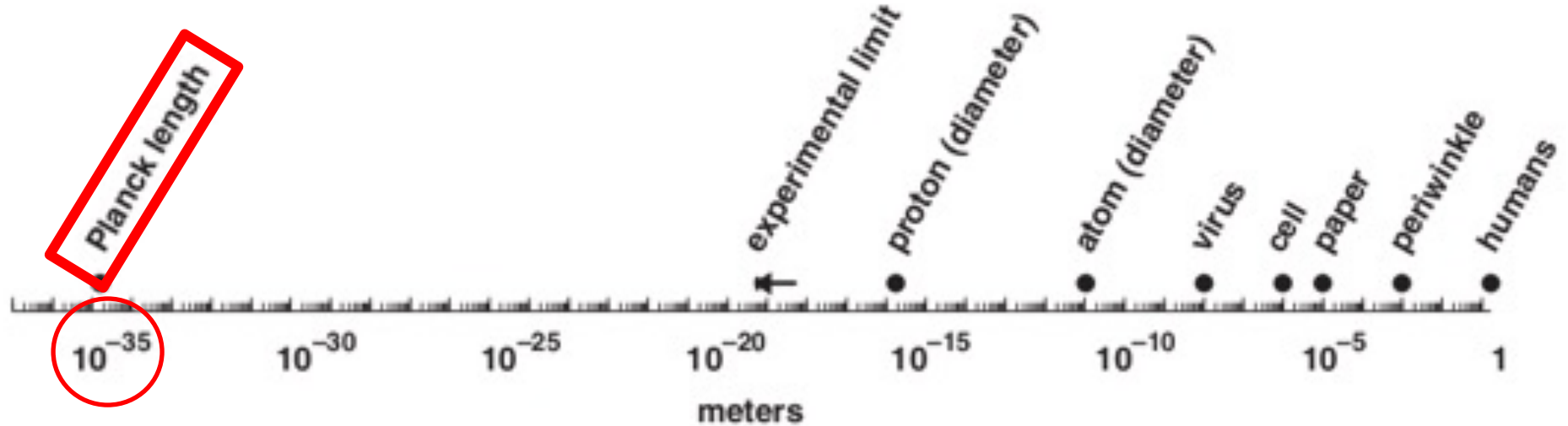
TIMOTHY PAUL SMITH

OXFORD

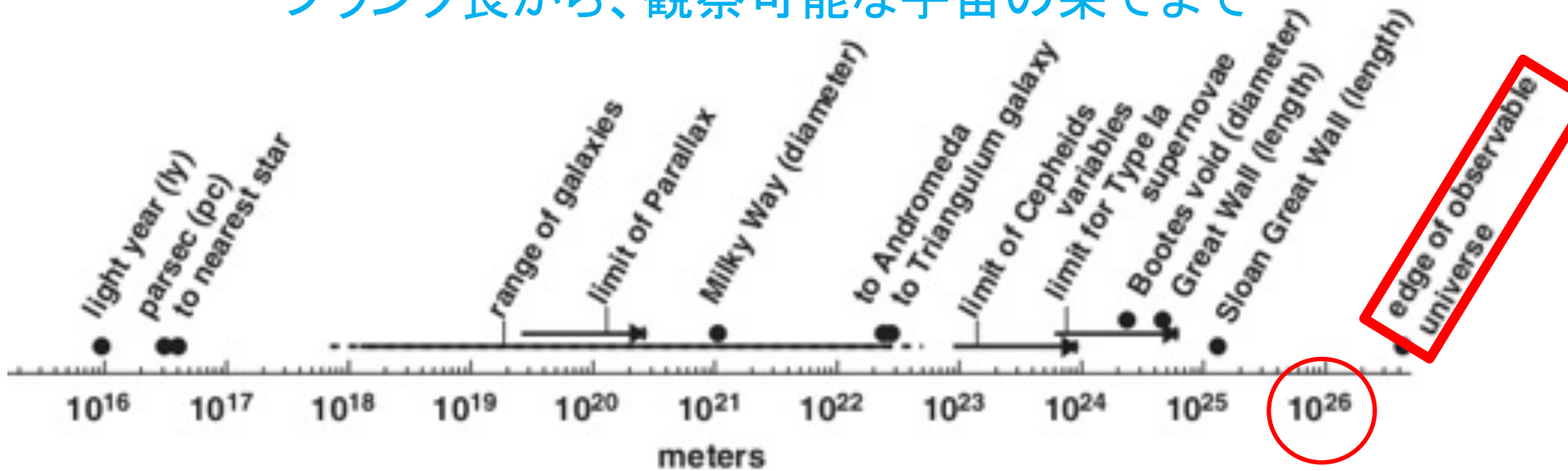
Copyrighted Material



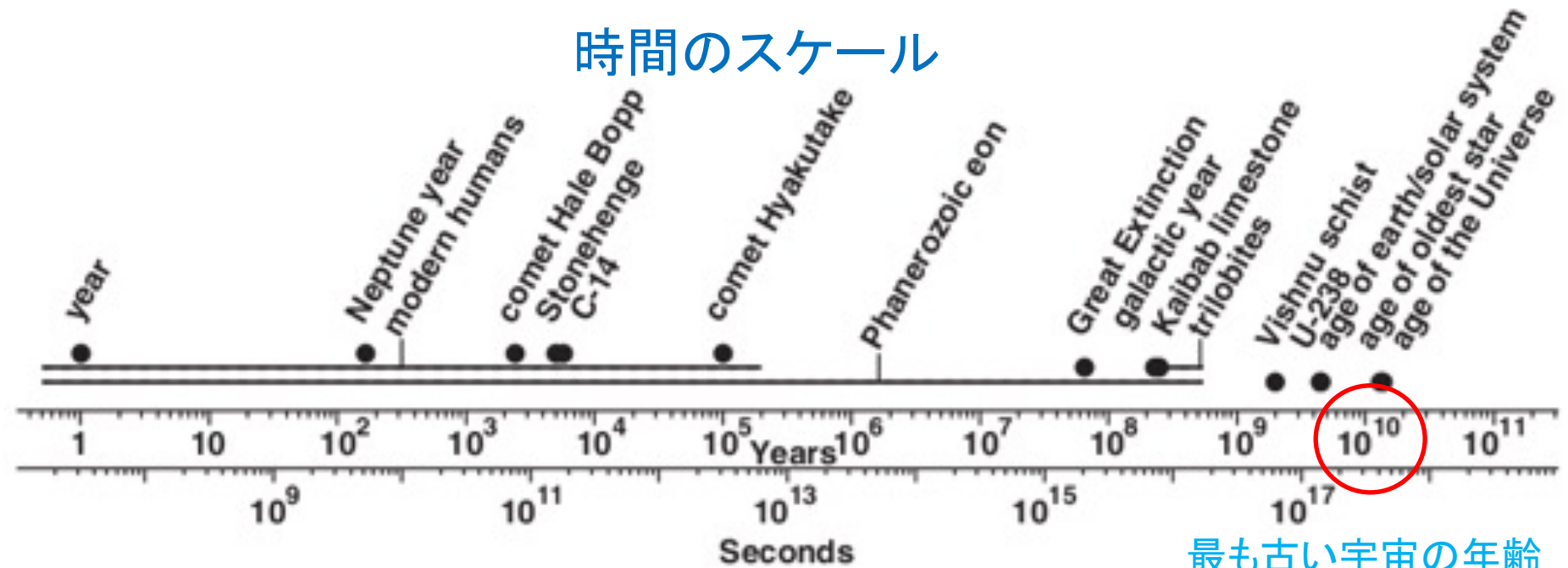
長さのスケール



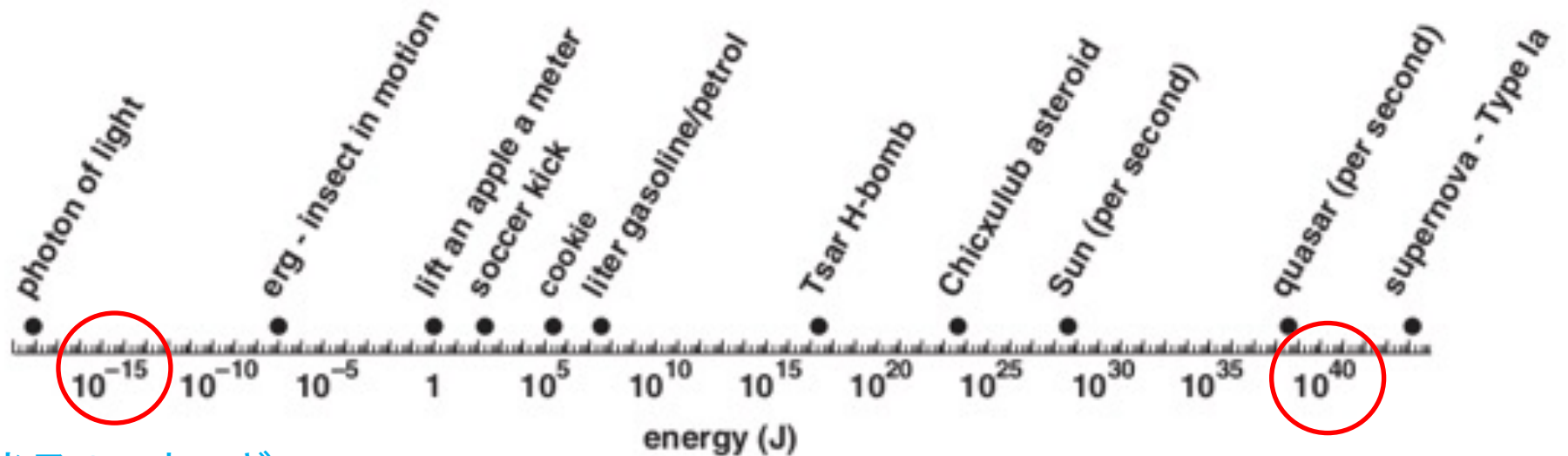
プランク長から、観察可能な宇宙の果てまで



時間のスケール



エネルギーのスケール



猿のタイピスト



猿のタイピスト

- チンパンジーにタイプライターを叩かせる。ずっとやっていれば、"monkey"という単語を打ち出す可能性はある。もっともっとやり続ければ、シェクスピア 작품을打ち出す可能性だってあるはずだ。
 - アルファベット(小文字)は、'a'から'z'まで26文字である。"monkey"は6文字だ。もしも猿が26キーのタイプライターを6回適当に叩いたとする。この時、猿がうちだすことが可能な文字列の総数は、26の6乗で、308,915,776になる。9桁の数字で3億ちょっとである。
-

宝くじに当たる確率と"monkey"と打ち出す確率

- 宝くじの番号は、75組 159149番(これが今年の5億円のあたりの番号)のように8桁なので、タイプライター猿が、
"monkey"と打ち出すのは、サマージャンボで5億円当てるより、30倍ほど難しいことがわかる。

それでも、誰かには5億円が当たるように、もしも、猿のタイプライターが3億匹いれば、そのうちの一匹は、"monkey"と打ち出す可能性は高いのだ。



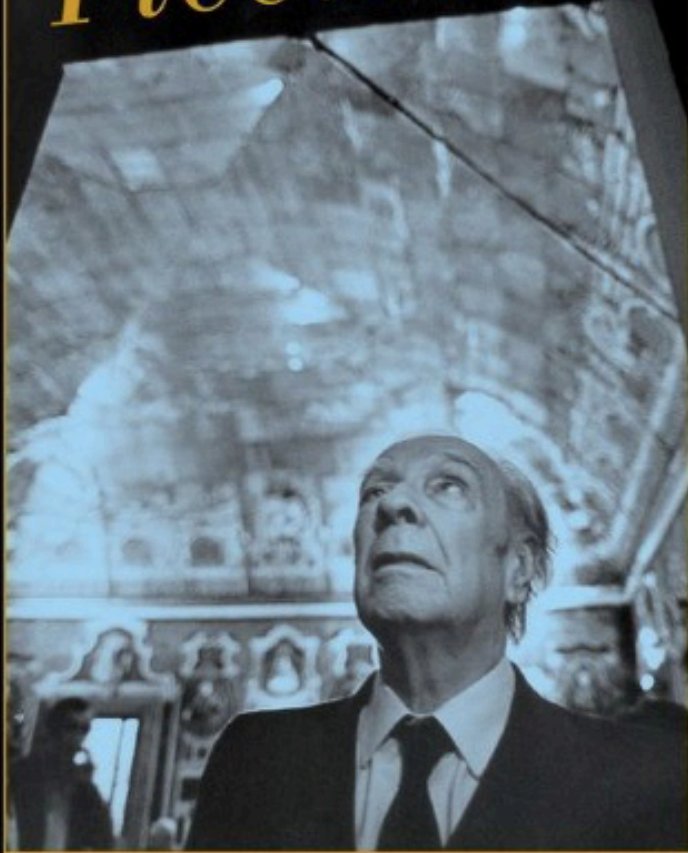
□ 「猿のタイピスト」については、日本語wikiの「無限の猿定理」
<https://goo.gl/3iM9V9> が、素晴らしいまとめになっている。
是非、読んでほしい。僕は、ボルヘスの「完全な図書館」の話
をしようと思っていたのだが、完全に負けた。特に、古代ギリ
シャの原子論に対する批判として、この議論が、すでにあった
ことは知らなかった。

□ どっちの見方が正しいのか？

□ 実は、この矛盾する見方が、交わるところこそ、熱力学的なエ
ントロピーと情報論的なエントロピーが交わるところになる。

"Without Borges the modern Latin American novel
simply would not exist." —CARLOS FUENTES

Ficciones



Jorge
Luis BORGES

" ... the Library is total and that its shelves contain all the possible combinations of the twenty-odd orthographic symbols (whose number, though vast, is not infinite); that is, everything which can be expressed, in all languages. Everything is there:

the faithful catalogue of the Library, thousands and thousands of false catalogues, a demonstration of the fallacy of these catalogues, a demonstration of the fallacy of the true catalogue"

Borges, Jorge Luis. "The Library of Babel" Ficciones

英語の単語の数

- ところで、英語の単語は、何個あるのだろうか？
 - 先に、"monkey"の例で、6文字のアルファベットの数、約3億 3×10^9 個あることがわかったので、12文字のアルファベットの並びの数は、約 9×10^{18} 、だいたい10の19乗ほどあることがわかる。随分大きな数だが、その全てが、意味のある単語にはなっていない。というか大部分が、猿タイプライターが打ち出すような無意味な文字の並びである。
 - (興味がある人は、「一番長い英単語は？」でググって見てほしい。)
-

単語ではなく文章を打ち出す

- ただ、「単語」ではなく「文章」を打ち出そうとすると、もっともっと大きな数が登場することになる。
- 無意味な綴りを乱発する猿のタイプライターは、クビにして、必ず意味のある単語を打ち出すが、単語の順序はデタラメであるタイプライター・フジオを雇うことにしよう。英語ではなく日本語で考えよう。文章を作るという点で、フジオは猿よりマシだろうか？
- 日本語の単語の数を10万としよう。
現実には、日本語の語彙の数は、もっと多い。

日本語の辞書の収録語彙数をあげておく。

『日本国語大辞典』(小学館) 50万語

『広辞苑』(第六版、岩波書店) 約24万語

『岩波国語辞典』(第七版) 6万5000語

新村 出編

広辞苑

第六版

岩波書店

新村 出編

広辞苑

第六版

岩波書店

新村 出編

広辞苑

第六版

岩波書店

10個の語からなる文の数

- 日本語で、10個の語からなる語の並びがどれくらいあるか数えてみよう。
- 10語文というのは、そんなに長い文章ではない。

古池₁ や₂ かわず₃ とびこむ₄ 水₅ の₆ おと₇
これが7語文。

この₁ 文₂ は₃ 10₄ 個₅ の₆ 語₇ から₈ できて₉ いる₁₀
これが、10語文の例である。

-
- ところが、語彙が10万個ある言語での10個の語からなる語の並びは、 $100000^{10} = 10^{50}$ 種類もある！ 10^{50} というのは、とても巨大な数である。
 - 文は語から構成され、語は文字（アルファベットだったり50音）で構成されているとすると、語の複雑さより、文の複雑さの方が、はるかに大きいのである。
-

ベンジオの次元の呪い

- そのことは、語彙については、収録範囲には違いがあるが、有限個の語彙を収めた「辞書」が存在するのに対して、ある言語で可能な表現を全て網羅した「用例集」が存在しないことを考えてみればわかる。
 - ボルヘスが夢想する「完全な図書館」は、存在し得ないのだ。完全な図書館が存在したとする。その時、その図書館の目録を考える。その目録を含めた目録が必要になる。いつまでたっても図書館は完全にはならない。
 - ディープラーニングの手法で自然言語を攻略しようとしていたベンジオは、文の複雑さの次元の大きさに、それを「次元の呪い」と呼んだ。
-

A microscopic image showing numerous cells, likely from a tissue sample. The cells are stained with a purple dye (likely hematoxylin) and a yellow/green dye (likely eosin). The purple staining highlights the nuclei of the cells, while the yellow/green staining highlights the cytoplasm and extracellular matrix. The cells are irregular in shape and size, and are densely packed in some areas.

19世紀 ボルツマン 統計力学とエントロピー

$$S = k \log W$$



LUDWIG
BOLTZMANN
1844 - 1906

DR. PHIL. PAULA
BOLTZMANN
GEB. CHLARI
1891 - 1977

$$S = k \log W$$



LUDWIG
BOLTZMANN
1844 - 1906

DR. PHIL. PAULA
BOLTZMANN
GEB. CHLARI
1891 - 1977

$$S = k \log W$$

- ボルツマンの墓には、彼が発見した、次のエントロピーの式が刻まれている。

$$S = k \log W$$

ここで、 S はエントロピーで、 k はボルツマンの定数、 W はミクロな状態の数である。

- 先に、なぜ、熱力学的には エネルギーを温度で割った単位を持つ量であるエントロピーが、情報の理論と結びつくのだろうかと書いたのだが、この式が鍵である。
-

$$S = k \log W$$

- このボルツマンの式は、エントロピーが、ミクロな状態の数の対数 $\log W$ に「比例」するという彼の偉大な発見を表しているのだが、情報の理論と結びつくのは、この $\log W$ の部分なのである。この部分は、単なる整数の対数なので、次元(単位のこと)は無い。
- 先の関係式で、「エネルギー 割る 温度」というエントロピーの単位(次元のこと)は、すべて、 k というボルツマンの定数の中に閉じ込められている。
- もしも、 $k = 1$ なら(実際、そういう表記をすることもある)

$$S = \log W$$

$$S = k \log W$$

- となって、エントロピーと E/T (Joul/Kelvin) という熱力学的単位との関係は、ほとんど見えなくなる。
- その意味では、ボルツマンの定数 k は、熱力学的エントロピーと情報論的エントロピーを結びつける、役割を持つのである。
(少なくとも、「単位・次元」を揃えるという意味では)
- スタンフォードの物理学講義で、サスキンドは、こんな冗談を言っていた。

「ボルツマンは、彼の定数を、誰も覚えてくれないので、それを悲しんで、自殺したんだ。」

原子論者ボルツマン

ボルツマンは、原子論者

- 物理現象を、より基本的な要素の存在とその運動で説明しようというのは、現代物理学の基本的な態度である。ボルツマンらがその基礎を築いた統計力学の手法は、20世紀の量子力学を準備する。素粒子論というのは、現代の原子論に他ならない。

フェルミオン			ボソン	
クォーク	u アップ	c チャーム	t トップ	γ 光子
	d ダウン	s ストレンジ	b ボトム	g グルーオン
レプトン	ν_e 電子ν	ν_μ ミューν	ν_τ タウν	W W ボソン
	e 電子	μ ミューオン	τ タウ	Z Z ボソン
				H ヒッグス

現代物理学の「標準モデル」での素粒子のリスト

異端の学説 原子論

- 古代ギリシャで原子論を提唱したデモクリトスの著作を、プラトンは、全部、焼かせたという。
- 「しかし、そのこと[粒子の偶然の衝突で宇宙が誕生したということ]が起こりえたと考える者は、どうして次のことに考えが至らないのか、わたしは理解に苦しむ。すなわち、黄金製であれ何であれ、**二十一種類のアルファベットの文字を数えきれないほど集めて何かある容器の中に投げ込み、それらを攪拌して地面に投げ出すと、たとえばエンニウスの『年代記』のように、読者にとってちゃんと読める形になって並ぶとはどうして考えないのか。**もっとも、わたしは、幸運の助けを借りたとしても、たった一行の詩句さえまともにできるかどうか、いぶかしく思っているのだが。」これは、3世紀のキケロの原子論批判。
- これは、「~~猿のタイプライター~~」の論理だ！

<https://ja.wikipedia.org/wiki/無限の猿定理>

Scott Aaronsonは、自書
“Quantum Computing
Since Democritus”
「デモクリトス以来の量子
コンピューティング」の表紙を
デモクリトスで飾った

QUANTUM COMPUTING SINCE DEMOCRITUS



SCOTT AARONSON

19世紀でも、原子の存在は、疑われていた

- ボルツマンは、日本でいえば、江戸時代末期に生まれ、明治維新の頃に論文を書いていると思っていい。
 - ただ、原子論が、科学の世界で広いコンセンサスを得るのは（化学の世界では、原子論は、早くから受け入れられていたようなのだが）、1905年の、アインシュタインの「ブラウン運動」についての論文以降だった。
 - 19世紀の物理学でも、「原子論」は「異端」の学説だったらしい。
-

ボルツマンの悲劇

- 19世紀の物理学者の双璧は、ボルツマンとマックスウェルだと思うが、二人のアカデミーでの人生は、はっきりと明暗を分けている。
- 紹介したサスキンドの発言は「悪い冗談」なのだが、ボルツマンの業績は、彼の生前には正當に評価されなかった。特に、当時の物理学会で大きな発言力を持っていたマッハとその取り巻きは、執拗にボルツマンの「原子論」を攻撃した。ボルツマンは、追い詰められ、次第に精神を病んでゆく。
- アインシュタインの画期的な論文「熱の分子論から要求される静止液体中の懸濁粒子の運動について」が出たのは、1905年5月である。ボルツマンも、この20代の若者の論文を読んだと思うのだが。
- 翌1906年、彼は、自ら命を絶つ。

ミクロの世界とマクロな世界 隠される情報と新しい情報の「出現」

みつけのねえとんざい人ぞ

大せいの
人々うく
たうく
いんと
とく人のと
い

い

一
國

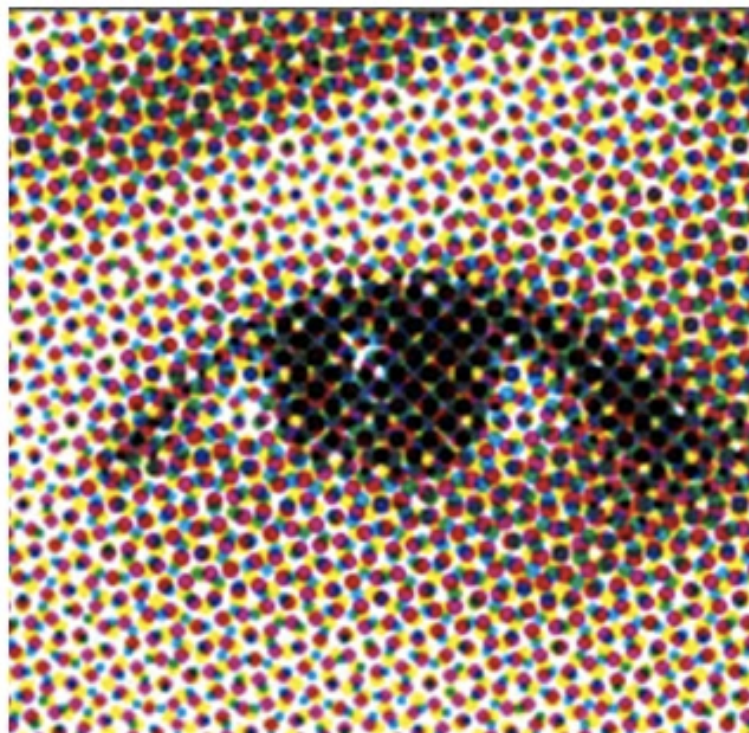
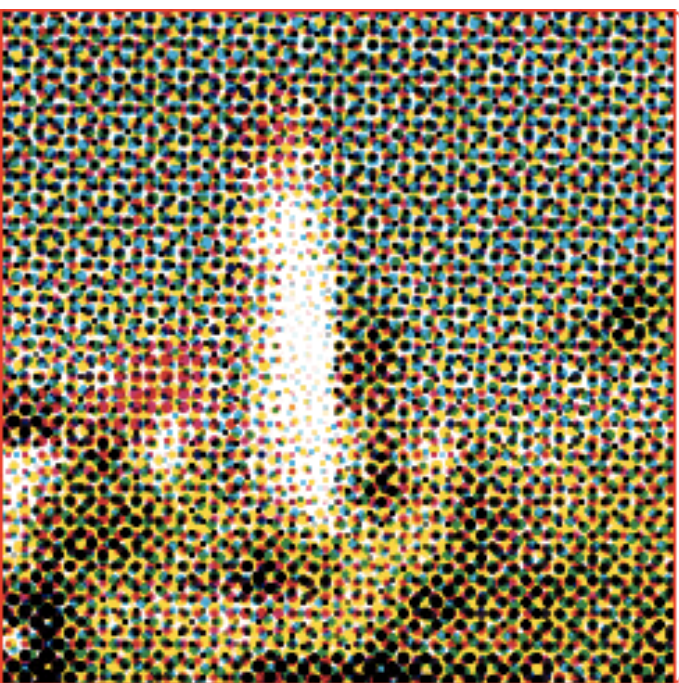
戲

751

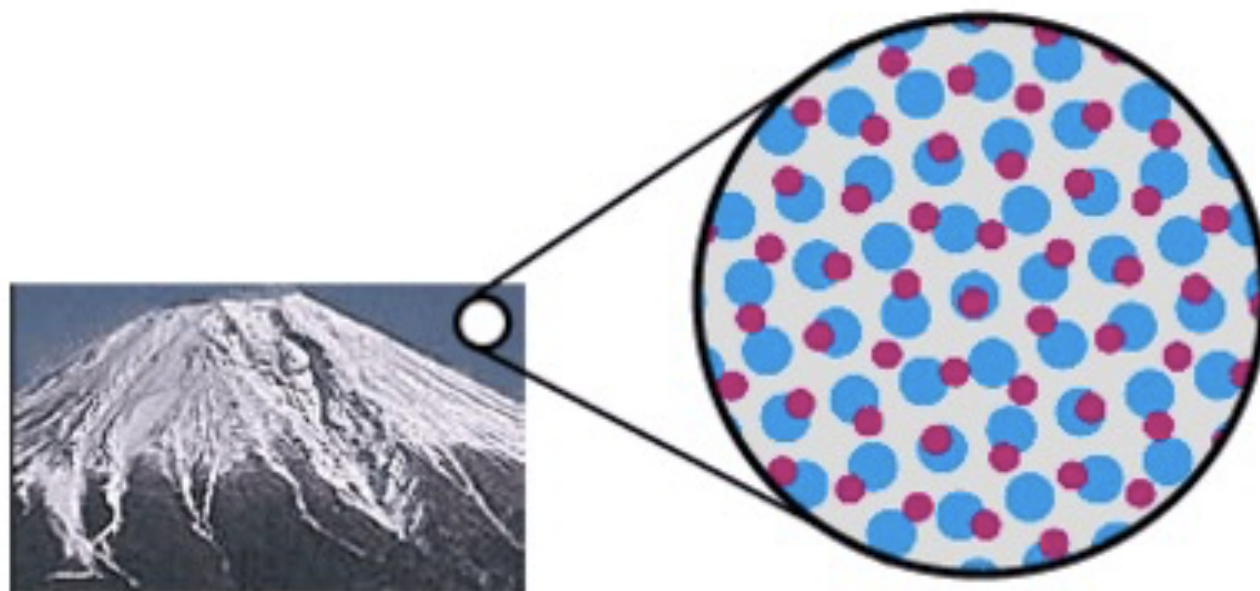
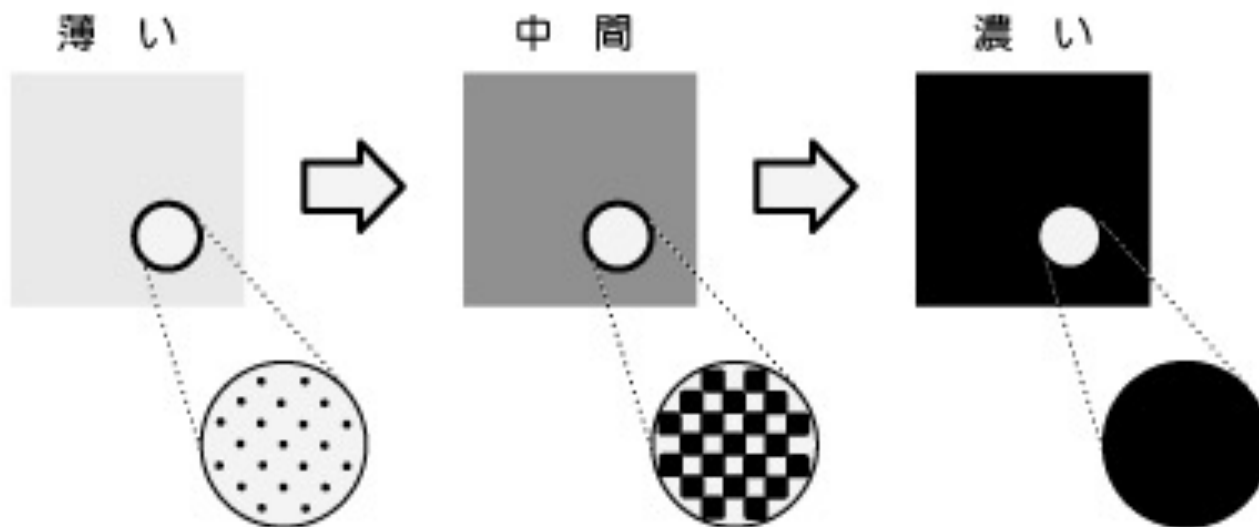








<https://goo.gl/jRmTA2>



情報の損失としてのエントロピー

John C. Baez, Tobias Fritz, and Tom Leinster:
"A Characterization of Entropy in Terms of
Information Loss"

<http://goo.gl/p13C4d>

「情報の損失」としてのエントロピーの特徴づけ

- 我々の新しい方法は、エントロピーそのものではなく、その変化に注目する。これは、多くの点で意味がある。
- あるプロセス p を経ることで、失われる情報を $F(p)$ で表す。

- $$F(f \circ g) = F(f) + F(g)$$

二つの段階からなるプロセスを考える。プロセス全体での情報の損失は、それぞれの段階での情報の損失を足したものだ。

「情報の損失」としてのエントロピーの特徴づけ

□
$$F(\lambda f \oplus (1 - \lambda)g) = \lambda F(f) + (1 - \lambda)F(g)$$

確率 λ のコインで、 f と g のどちらのプロセスを選ぶか決めるとする。この時の情報の損失は、 f を選んだ時に失う情報に λ を掛け、それに、 g を選んだ時に失われる情報に $(1 - \lambda)$ をかけたものを足したものである。

□ 連続性:

プロセスをほんの少し変えれば、情報の損失も、ほんの少しだけ変わる。

□
$$F(f) = c(S(p) - S(q))$$

情報の損失は、シャノンのエントロピーの変化に比例する。

「情報の損失」としてのエントロピーの特徴づけ

- こうした前提から、自然に、次の式を導くことができる。

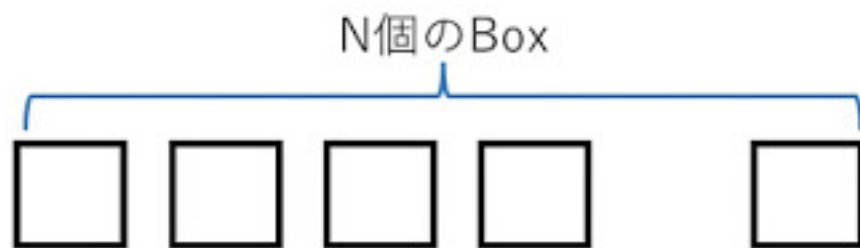
$$S(p) = - \sum_{i \in X} p_i \ln(p_i)$$

- あとで見るように、シャノン は、エントロピーの式を、5つの条件から、それらを満たす唯一の式として数学的に導いている。
ここでのバエズの導出は、それより簡単であるだけでなく、それぞれの条件の意味も、わかりやすい。
- 「エントロピーそのものより、その変化が重要」
それは、「情報の損失」として理解できる。
~~このことはまた、「相対エントロピー」の重要性も、示唆している。~~

$S = k \log W$ を計算してみよう！

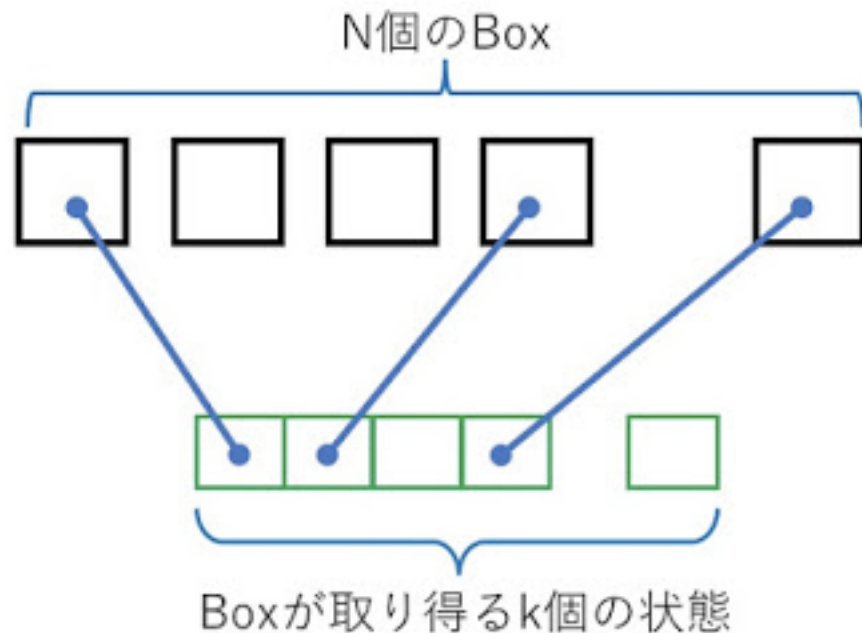
ある系の持つ可能な状態の数を考える

- まず、同じ大きさのN個のBoxからなる系を考えよう。それぞれのBoxはk個の状態を持つとしよう。

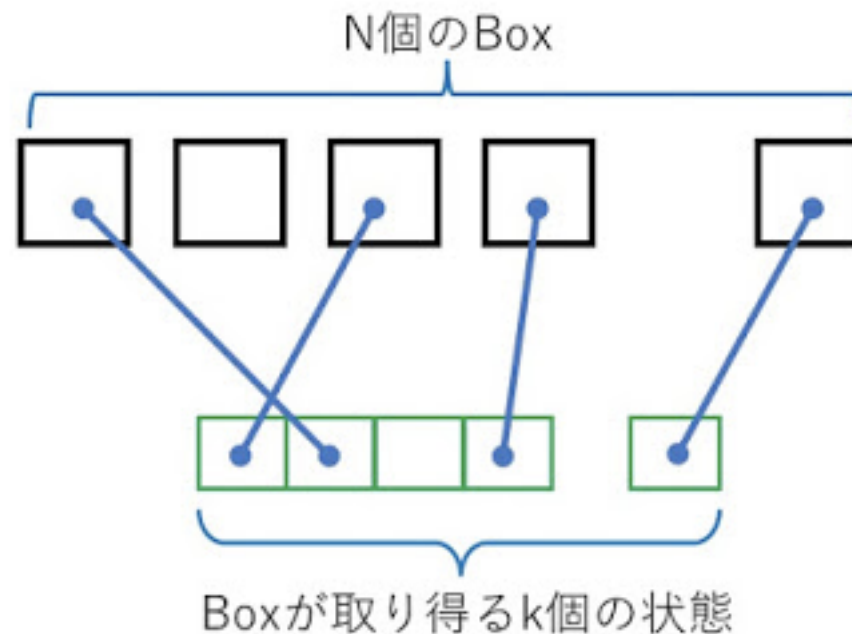


Boxが取り得るk個の状態

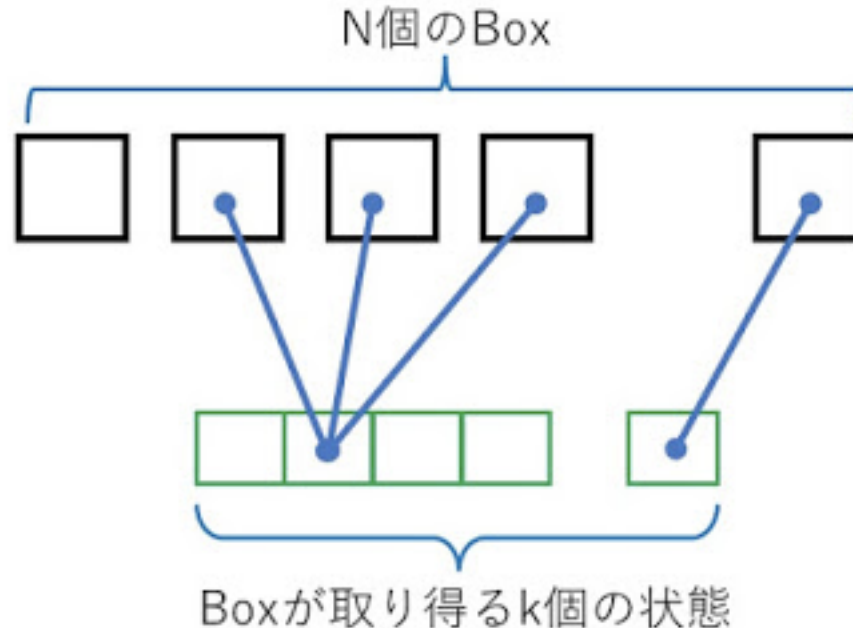
- あるBoxがある状態*i*を取るとき、次のように線で結ぶ。
- この例は、Box1が状態1を取り、Box4が状態2を取り、Box Nが状態4を取ることを表している。Box2, Box3は、この図では線で状態と結ばれていないが、それは、表記上省略しているだけである。



- 次の図では、Box1が状態2を取り、Box3が状態1を取り、Box4が状態4を取り、Box Nが状態kを取ることを表している。Box2の状態は、この図では省略されている。



- 全てのBoxは、必ず、何らかの状態を持つのだが(例では省略されていても)、全ての状態が、あるBoxの状態になるとは限らない。次の例では、Box2, Box3, Box4 は、同じ状態2を持っている。



- 状態*i*が、何個のBoxで共有されているかを n_i で表すことにしよう。先の例では、 $n_2=3$ ということになる。
 $n_1+n_2+n_3+\cdots+n_k=N$ である
- この時、 N 個のBox、 k 個の状態からなるシステムが取り得る可能な組み合わせの総数 W は、次のようにして計算できる。

$$W = \frac{N!}{n_1!n_2!n_3\cdots n_k!}$$

log Wを計算してみる

- 階乗は、少し扱いにくいので、次のスターリングの近似式を用いる。

$$\log n! \simeq n \log n - n$$

$$\log W = \log \left(\frac{N!}{n_1! n_2! n_1 \cdots n_k!} \right) = \log N! - \sum \log n_i! :$$

$$= (N \log N - N) - \sum (n_i \log n_i - n_i)$$

- 先に見たように、 $\sum n_i = N$ なので、左の $-N$ と右の $\sum n_i$ は、打ち消しあって消える。

$$\log W = N \log N - \sum n_i \log n_i$$

$$\log W = N \log N - \sum n_i \log n_i$$

- ここで、 $p_i = \frac{n_i}{N}$ とおく。 $n_i = p_i N$ である。
また、 $\sum n_i = N$ なので、 $\sum p_i = 1$ である。

$$\begin{aligned} \frac{1}{N} \log W &= \log N - \frac{1}{N} \left(\sum n_i \log n_i \right) \\ &= \log N - \sum p_i \log(p_i N) \\ &= \log N - \left(\sum p_i \log p_i + \sum p_i \log N \right) \end{aligned}$$

- $\sum p_i = 1$ なので、 $\sum p_i \log N$ は、 $\log N$ となって打ち消しあう。

$$\frac{1}{N} \log W = - \sum p_i \log p_i$$

分配関数 Partition Function

分配関数

- 先のBoxの図で、それぞれのBoxが状態*i*を持つときに、エネルギー E_i を持つとする。系全体のエネルギー $Energy_{Total}$ は、次のようになる。ここで、 $\langle E \rangle$ は、一つのBoxの平均エネルギーである。

$$Energy_{Total} = \sum n_i E_i = N \langle E \rangle$$

- 両辺を N で割って、 $p_i = \frac{n_i}{N}$ を使えば、次の式が成り立つことがわかる。

$$\sum p_i E_i = \langle E \rangle$$

- この時、 $-\sum p_i \log p_i$ は、どのような時に最大になるのか考えよう。

ラグランジェの未定乗数法

- 拘束条件が与えられた時の式の最大値(最小値でもいいのだが)の(必要)条件は、ラグランジェの未定乗数法で求められる。
- 次の式を考える。

$$\mathcal{L} = \left(-\sum_i p_i \log p_i \right) + \alpha \left(\sum_i p_i - 1 \right) + \beta \left(\sum_i p_i E_i - \langle E \rangle \right)$$

- この式を p_i で偏微分した式がゼロになる時、元の式が極値を取る条件が求められる。

$$0 = \frac{\partial \mathcal{L}}{\partial p_i} = -\log(p_i) - 1 + \alpha + \beta E_i$$

$$\log p_i = -1 + \alpha + \beta E_i; \quad p_i = \exp(-1 + \alpha + \beta E_i),$$

これから、 $\log p_i = -1 + \alpha + \beta E_i$ がわかるから、 $p_i = \exp(-1 + \alpha + \beta E_i)$ 。

この時、 $1 = \sum_i p_i = \exp(-1 + \alpha) Z$ と置ける。ただし、 $Z \equiv \sum_i \exp(\beta E_j)$ である。

だから、 $p_i = \frac{1}{Z} \exp(\beta E_i)$ すなわち、 $p_i = \frac{e^{\beta E_i}}{\sum_j e^{\beta E_j}}$

$$Z \equiv \sum_i \exp(\beta E_j)$$

の形の式を「(エネルギー)分配関数(partition function)」と呼び、多くの応用がある。

温度を T 、ボルツマン定数を k_B とした時、 $\beta \equiv \frac{1}{(k_B T)}$ であるので、熱力学では、次の形の分配関数がよく用いられる。

$$Z = \sum_j \exp\left(\frac{E_j}{k_B T}\right)$$

もう一度、式 $S = k \log W$ を考える

離散的な量としてのエントロピー

□ ボルツマンのエントロピーの式

$$S = k \log W$$

を見ると、あることに気づく。

- W は、ミクロな状態の「数」だというのだから、 $\log W$ は、ある整数の対数を取ったもの。 $W=1$ の時、 $\log 1 = 0$ だから $S = 0$ となるが、次の値は、 $W=2$ の時で、 $S = k \log 2$ 。その次は、 $S = k \log 3$ である。この式で表されるエントロピーの値は、決して任意の連続な値をとるわけではない。
-

離散的な量としてのエントロピー

- 現在の目から見れば、この式は、エントロピーの離散的で「量子論的」性質を表していると解釈できるのだが、ボルツマンの生きた19世紀には、量子論はまだなかった。量子論は、彼が、マッハらに攻撃された「原子論」の、自然な帰結なのだということも、よくわかる。
 - 同時に、この時代には、エントロピーの離散的な解釈を与える、シャノンの情報理論も存在していない。不思議な関係式だ。
-

フォン・ノイマンによる エントロピーの量子論的定式化

- エントロピーの量子論的定式化を行ったのは、フォン・ノイマンである。もちろん、シャノンの登場のはるか以前である。驚くほどの慧眼である。
- すでに、1932年の「量子力学の数学的基礎」の中で、次のようなエントロピーの定式化を与えている。

$$S = - \text{Tr} (\rho \log \rho)$$

ここで、 ρ は密度行列、 Tr は行列のTrace(対角要素の和をとったもの)である。この式の説明は、別の機会にゆずるが、ギブスの与えた密度関数 ρ を用いたエントロピーの式と、基本的には同じ形をしている。(ΣもTrも、和をとる演算である)

$$S = - \sum (\rho \log \rho)$$

フォン・ノイマンのもう一つの「貢献」

- エントロピーの歴史の中で、フォン・ノイマンは、エントロピーの量子化だけでなく、もう一つの重要な「貢献」をしている。
- 1940年、それまでの熱力学的エントロピー概念の系譜とは全く独立にシャノンが発見した「不確実さの関数」＝「情報量」の概念に、「エントロピー」という名前をつけることを勧めたのは、フォン・ノイマンだと言われている。<https://goo.gl/Lkm98X>
- “You should call it ‘entropy’ for two reasons: First, the function is already used in thermodynamics under that name; second, and more importantly, most people don’t know what entropy really is, and if you use the word ‘entropy’ in an argument you will win ever time.”
- “nobody really knows what entropy is anyway.”
「誰も、エントロピーがなんなのかなんて知っちゃいない」



The background of the slide is a light lavender color, populated with numerous irregular, organic shapes. These shapes are primarily composed of small, overlapping circles and dots in shades of deep purple, magenta, and a metallic gold or bronze. The distribution of these shapes is somewhat random but creates a textured, almost cellular or molecular appearance across the entire frame.

20世紀 シヤノン 情報理論とエントロピー

シャノン、エントロピーを再定義する

A Mathematical Theory of Communication

C. E. SHANNON

<http://www.qsl.net/n9zia/pdf/shannon1948.pdf>

1948年

目標: あるメッセージに関連づいた情報の量の尺度を構成すること

あるメッセージを受け取った時に得られる情報の量は、そのメッセージがどの程度ありそうなものに依存する。

あるメッセージが、ありそうでないようなものであればあるほど、それを受け取った時の情報は、ますます大きくなる。

Def. 1: $X = \{x_1, x_2, \dots, x_l\}$ を、 l 個のメッセージの集合としよう。 X 上の確率分布 $P = (p_1, p_2, \dots, p_l)$ とは、それぞれのメッセージ x_j に対して、確率 $p_j = p(x_j)$ を割り当てたものである。

このことは、 $p_j \geq 0$ and $p_1 + p_2 + \dots + p_l = 1$ を意味することに注意。

情報の尺度が満たすべき条件

Def. 2: X についての情報の尺度とは、次の条件を満たす実数値を取る関数 $H(X)$ である。 $H(X) : \{X \text{ の確率分布} \} \rightarrow \mathbb{R}$

- **Continuity**. $H(p_1, \dots, p_\ell)$ は、連続である。
- **Additivity**. 確率分布 P, Q に対して、 $H(p_1 q_1, \dots, p_\ell q_\ell) = H(P) + H(Q)$
- **Monotonicity**. 一様な確率分布 P, Q に対しては、 ℓ につれて、情報は増大する。 $P = (1/\ell, \dots, 1/\ell)$ で $Q = (1/m, \dots, 1/m)$ の時、 $m > \ell$ なら $H(Q) > H(P)$
- **Branching**. $H(p_1, \dots, p_\ell)$ は、そのプロセスが、どのように部分に分割されても独立である。
- **Bit normalization**. 二つの同じ確率を持つメッセージから取得される情報は、一ビットである。 $\therefore H(1/2, 1/2) = 1$

この定義はどのような意味で情報の尺度になるか？

- Claim (Shannon 1949): この基準を満たす関数は、正確に一種類しか存在しない。すなわち、次のシャノンのエントロピー(シャノンの情報量)である。

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- $H(X)$ $p_1 = p_2 = \dots = p_{\ell} = 1/\ell$ の時最大になる。
- $H(X) = 0$ となるのは、一つの p_j のみが 1 で、その残りは、すべて0の場合である。
- $\log$ の底は2である : $\log_2 x = y \Rightarrow x = 2^y$.
- Bit normalization requires:
- If $X = \{x_1, x_2\}$, and $P = (1/2, 1/2)$, then $H(X) = 1$.
 - $H(X) = - (1/2 \log 1/2 + 1/2 \log 1/2) = \log 2$.
 - $\log 2 = 1$ if and only if $\log$ is to base 2.

不確かさの尺度としての $H(X)$

$P = (p_1, \dots, p_\ell)$ を、ランダムな変数上に値をとる

$X = \{x_1, \dots, x_\ell\}$ の確率分布とする。

□ Def. 1. X の期待値 $E(X)$ は次の式であたえられる

$$E(X) = \sum_{j=1}^{\ell} p_j x_j$$

□ Def. 2. もし、 X が値 x_j をとると測定された時、獲得された情報は、 $-\log_2 p_j$ であたえられる

■ 動機: p_j が大きいほど、 x_j は確実なものになり、それに関連づけられた情報は、より小さいものになるべきである。

獲得された情報 $-\log_2 p_j$ の期待値は、シャノンの情報量になる

$$E(-\log_2 p_j) = -\sum_{j=1}^{\ell} p_j \log_2 p_j = H(X)$$

$H(X)$ は、測定 X 上で期待される情報の獲得を示す

確率 $p(x_i)$ は、どのように解釈されるべきか？

□ 不確実さを強調する

- メッセージ x_i の情報のコンテンツは、受け取り手にとっては、それがいかに不確実かである。
- それゆえ、その確率は、認識論的なものである。
- 特に、 $p(x_i)$ は、メッセージ x_i の正確さについての、受け取り手の信念の程度の尺度である。

□ しかし、この確率は、ソースの性質によって設定されたものだ

- もし、そのソースが、確率論的なものでなかったら、 $p(x_i)$ は、認識論的に解釈できる。
 - もし、ソースが、もともと確率論的なものだとする、 $p(x_i)$ は、ソースがメッセージ x_i を生成する存在論的確率と解釈できる。
-

メッセージの圧縮の最大量としての $H(X)$

メッセージの圧縮の最大量としての $H(X)$

- $X = \{x_1, \dots, x_l\}$ を、それからメッセージが構成される文字の集合としよう
 - メッセージは、 N 文字からなるとする。
 - 確率分布 $P = (p_1, \dots, p_l)$ は、文字集合上のものになる。
その意味は
 - それぞれの文字 x_i が、メッセージで生起する確率を p_i とする
 - 別の言い方をすれば、典型的なメッセージは、文字 x_1 を、 $p_1 N$ 含み、文字 x_2 を、 $p_2 N$ 含む。
-

英文テキストでのアルファベットの出現確率

i	a_i	p_i							
1	a	0.0575	a	■	16	p	0.0192	p	■
2	b	0.0128	b	■	17	q	0.0008	q	■
3	c	0.0263	c	■	18	r	0.0508	r	■
4	d	0.0285	d	■	19	s	0.0567	s	■
5	e	0.0913	e	■	20	t	0.0706	t	■
6	f	0.0173	f	■	21	u	0.0334	u	■
7	g	0.0133	g	■	22	v	0.0069	v	■
8	h	0.0313	h	■	23	w	0.0119	w	■
9	i	0.0599	i	■	24	x	0.0073	x	■
10	j	0.0006	j	■	25	y	0.0164	y	■
11	k	0.0084	k	■	26	z	0.0007	z	■
12	l	0.0335	l	■	27	—	0.1928	—	■
13	m	0.0235	m	■					
14	n	0.0596	n	■					
15	o	0.0689	o	■					

この時、

典型的な、異なる
メッセージの数

$$= \frac{N!}{(p_1 N)! (p_2 N)! \cdots (p_\ell N)!}$$

だから

$$\log_2 \left(\text{典型的な、異なる
メッセージの数} \right) = \log_2 \left(\frac{N!}{(p_1 N)! (p_2 N)! \cdots (p_\ell N)!} \right)$$

$$\log_2 \left(\frac{N!}{(p_1 N)! (p_2 N)! \cdots (p_\ell N)!} \right) = \log_2(N!) - \{ \log_2((p_1 N)!) + \cdots + \log_2((p_\ell N)!) \}$$

スターリングの公式: $\log(N!) = N \log N - N$ を使う

$$\approx (N \log_2 N - N) - \{ (p_1 N \log_2 p_1 N - p_1 N) + \cdots + (p_\ell N \log_2 p_\ell N - p_\ell N) \}$$

$$= N \{ \log_2 N - 1 - p_1 \log_2 p_1 - \underline{p_1 \log_2 N} + \underline{p_1} - \cdots - p_\ell \log_2 p_\ell - \underline{p_\ell \log_2 N} + \underline{p_\ell} \}$$

$$= -N \sum_{j=1}^{\ell} p_j \log_2 p_j$$

$$\begin{aligned} & - (p_1 \log N + p_2 \log N + \cdots + p_i \log N) \\ & = - (p_1 + p_2 + \cdots + p_i) \log N = \log N \\ & \mathbf{p_1 + p_2 + \cdots + p_i = 1} \end{aligned}$$

$$= NH(X)$$

ボルツマンとシャノンの エントロピーの比較

ボルツマン

- ガス
- 異なるミクロな状態を数え上げる
- N =ミクロな一つの粒子の数
- $\{n_1, \dots, n_l\}$ = l -cellへの分布
- $\{p_1, \dots, p_l\}$ =cell上の確率分布
- $p_i = n_i/N$ =ミクロな状態がcellにある確率
- Np_i =cell内のミクロ状態の数

$$S_B(\Gamma_{D_i}) = -Nk \sum_{j=1}^{\ell} p_j \ln p_j$$

$$\frac{N!}{(p_1 N)!(p_2 N)! \cdots (p_{\ell} N)!}$$

シャノン

- メッセージ
- 異なるメッセージを数え上げる
- N =メッセージ中の文字数
- $\{x_1, \dots, x_l\}$ = l 文字のアルファベット
- $\{p_1, \dots, p_l\}$ =文字上の確率分布
- $p_i = n_i/N = x_i$ がメッセージにある確率
- Np_i =メッセージ内の x_i の数

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

シャノンの情報量/エントロピーと 他のエントロピー概念との関係

Thermodynamic entropy: $\Delta S = S_f - S_i = \int_i^f \frac{\delta Q_R}{T}$

Boltzmann entropy: $S_B(\Gamma_{D_i}) = -Nk \sum_{j=1}^{\ell} p_j \ln p_j + \text{const.}$

Gibbs entropy: $S_G(\rho) = -k \int_{\Omega} \rho(x, t) \ln(\rho(x, t)) dx$

Shannon entropy: $H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$

この時、

$$\log_2 \left[\begin{array}{c} \text{典型的な、異なる} \\ \text{メッセージの数} \end{array} \right] = NH(X)$$

だから

$$\left[\begin{array}{c} \text{典型的な、異なる} \\ \text{メッセージの数} \end{array} \right] = 2^{NH(X)}$$

だから、N文字のメッセージは、 $2^{NH(X)}$ 個しか存在しない。
このことは、メッセージのレベルでは、 $2^{NH(X)}$ bit を使って
エンコードするしかないということ。

2 possible messages require 1 bit: **0, 1**.

4 possible messages require 2 bits: **00, 01, 10, 11**.

それでは、文字レベルでは、n文字のアルファベットから、
N文字のメッセージをエンコードするのに、何 bit 必要に
なるか？

第一：n文字のアルファベットをエンコードするのに何bit 必要か？

- $n = \text{文字数}$ $x = \text{文字ごとのbit数}$
 - 2 文字 1 bit: 0, 1
 - 4 文字 2 bits: 00, 01, 10, 11
 - 8 文字 3 bits: 000, 001, 010, 011, 100, 101, 110, 111
 - だから: $n = 2^x$, so $x = \log_2 n$.
 - 1文字 $\log_2 n$ bits ということは、N文字長のシーケンスについては、 $N \log_2 n$ bits ということ。
 - ところで、メッセージ・レベルでは、 $NH(X)$ bitsでエンコードできた。だから、メッセージをエンコードするのに、 $N \log_2 n$ bitsは必要ではなく、 $NH(X)$ bitsで可能である。
 - こうして、 $H(X)$ は、与えられた文字から作られたメッセージが、圧縮されうる最大量を与える。
-

- $X = \{A, B, C, D\}$ とする ($n = 4$)
 この時、1文字につき $\log_2 4 = 2$ bits が必要である
 例えば、: $A = 00, B = 01, C = 10, D = 11$.
- **N文字のメッセージをエンコードするのに、 $2N$ bits が必要**
- それぞれの文字が、典型的なN文字のメッセージに出現する確率が、次のようだとしよう。
 $p_A = 1/2, p_B = 1/4, p_C = p_D = 1/8$
- この時、こうした典型的なN文字のメッセージの数は、

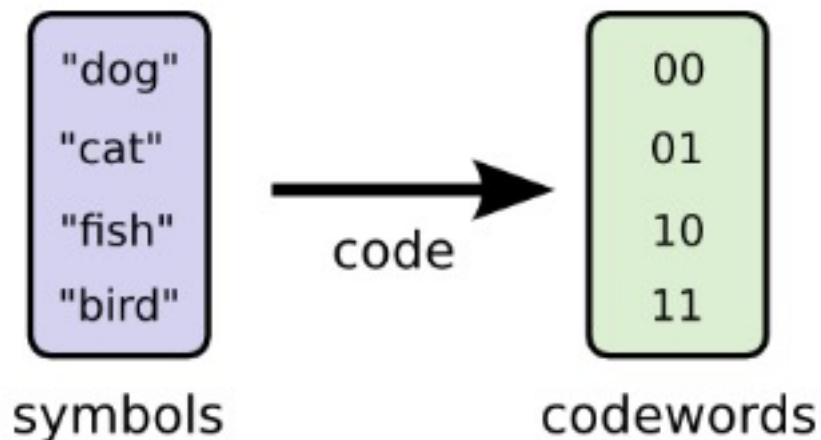
$$NH(X) = -N \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{8} \log_2 \frac{1}{8} \right) = 1.75N$$

この時、N文字のメッセージをエンコードするのに、 $2N$ bits ではなく、 $1.75N$ bitsでいいことになる。

- 全ての文字が等しい確率で出現するとすると

$$NH(X) = -N \left(\frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4} \right) = 2N.$$

固定長のコード 4 words 2bit



0 0 0 1 0 0 1 1 encoded string

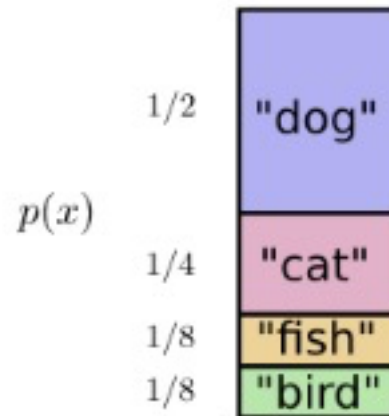
00 01 00 11 codewords

dog cat dog bird source symbols

Visual Information Theory

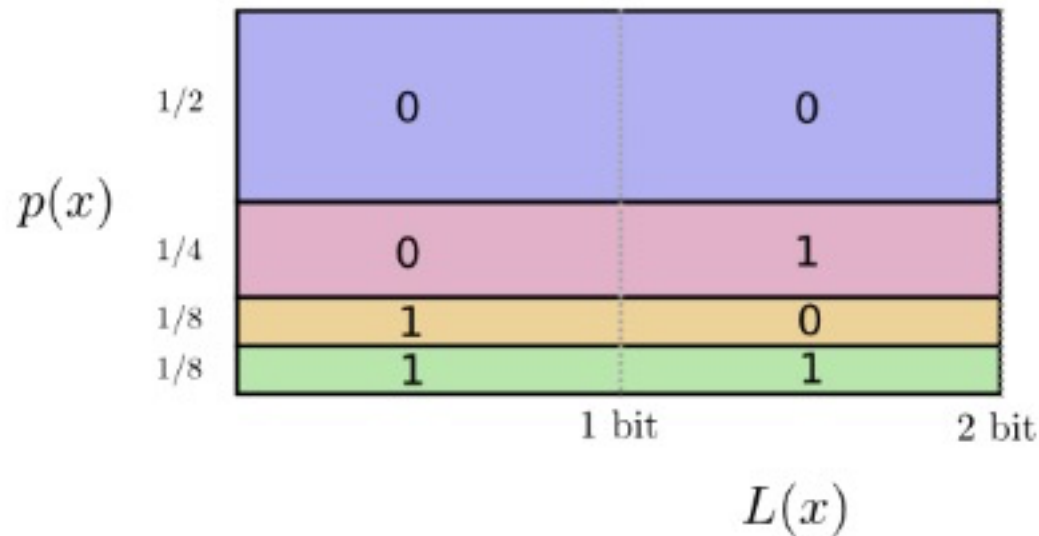
<http://colah.github.io/posts/2015-09-Visual-Information/>

語の出現頻度に 違いがある

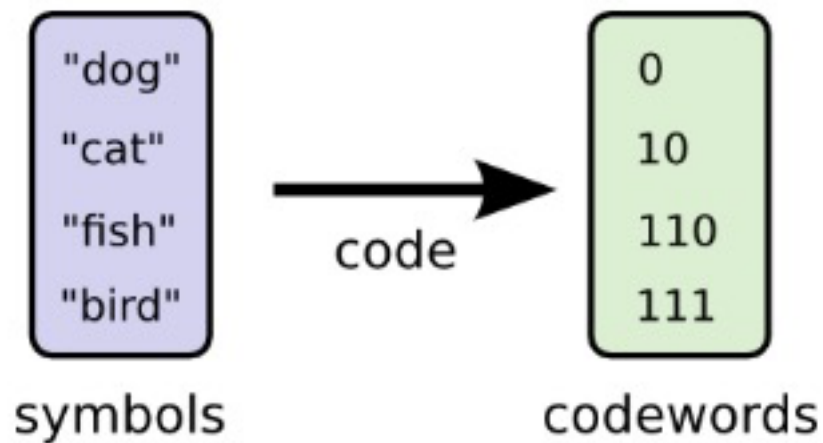


犬好きな人の
語の出現頻度

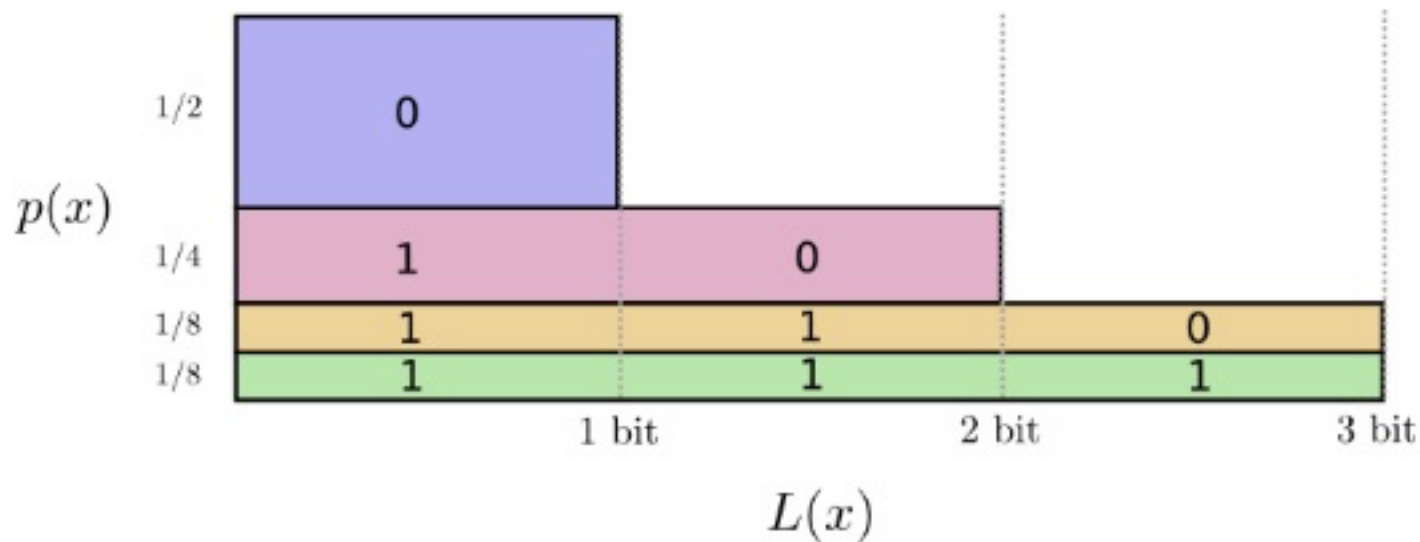
Old Code
2 bits



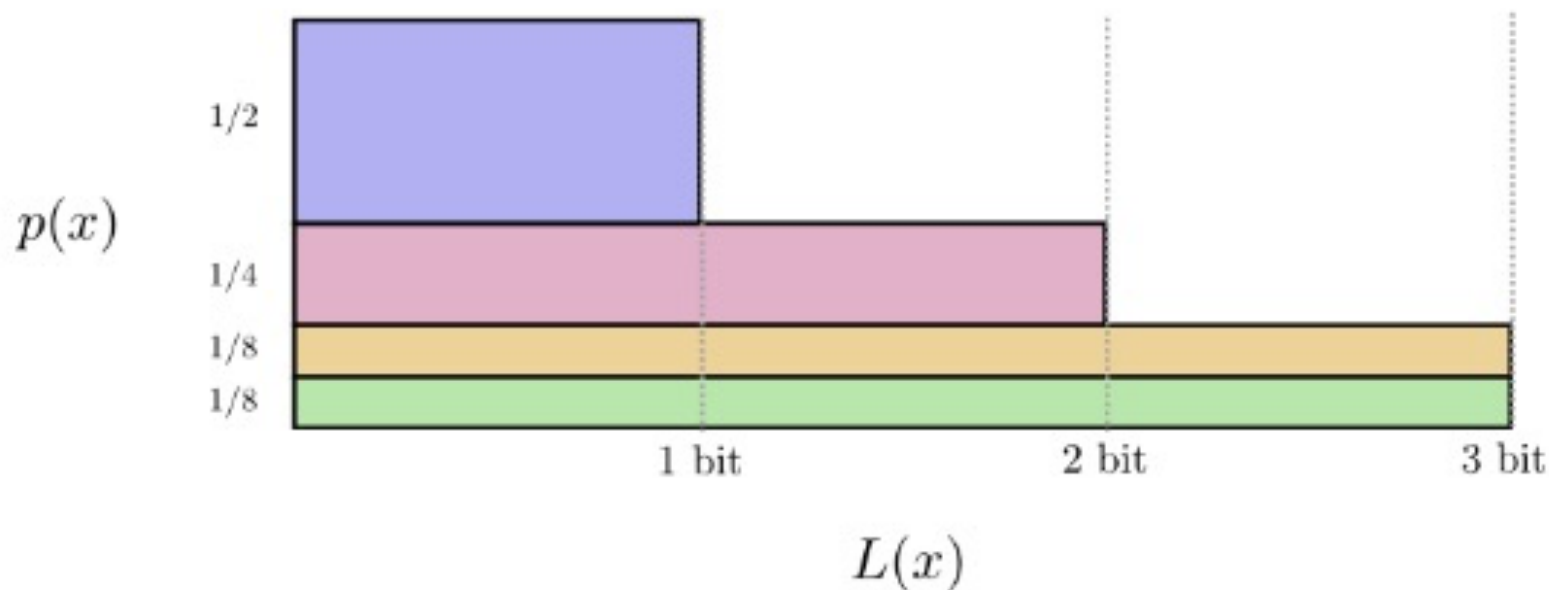
可変長コード

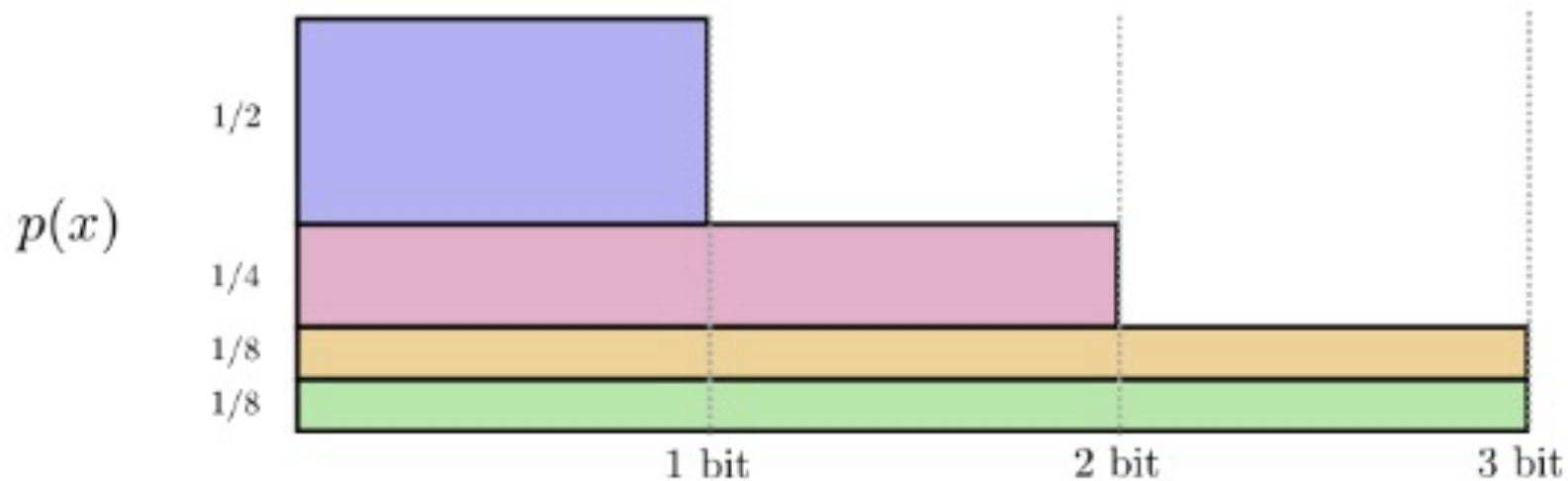


New Code 1.75 bits



エントロピー = 最適なコードの平均長
= 下の図の面積





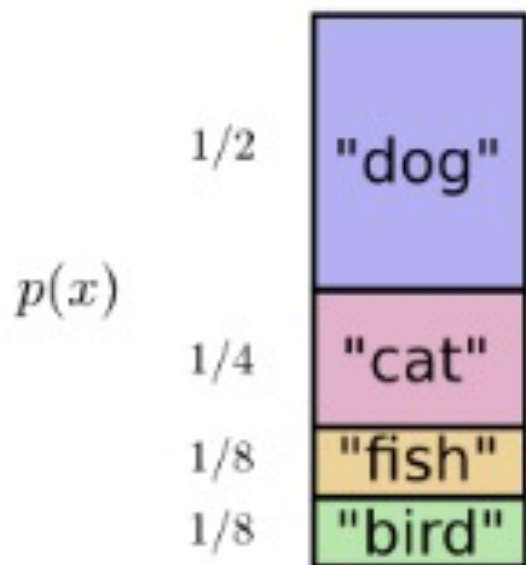
$$L(x) = \log_2 \left(\frac{1}{p(x)} \right)$$

エントロピー を計算する

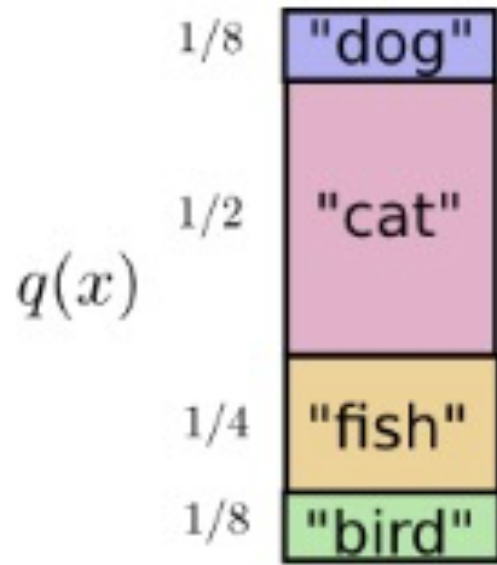
$$H(p) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right)$$

コード圧縮で Cross Entoropyを考える

Cross-Entropy



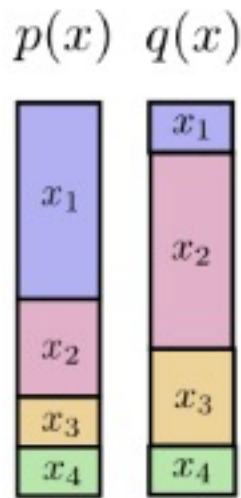
犬好きの人の
語の出現頻度
Bob



猫好きの人の
語の出現頻度
Alice

Cross-Entropy

$$H_p(q) = \sum_x q(x) \log_2 \left(\frac{1}{p(x)} \right)$$

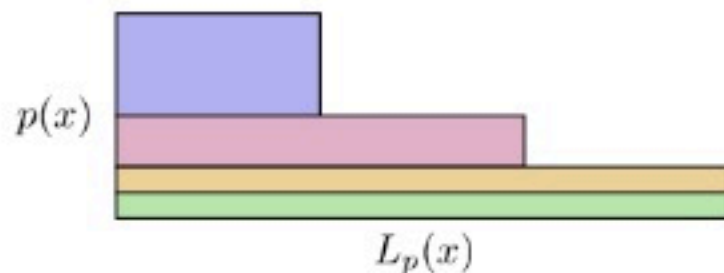


Cross-Entropy: $H_p(q)$

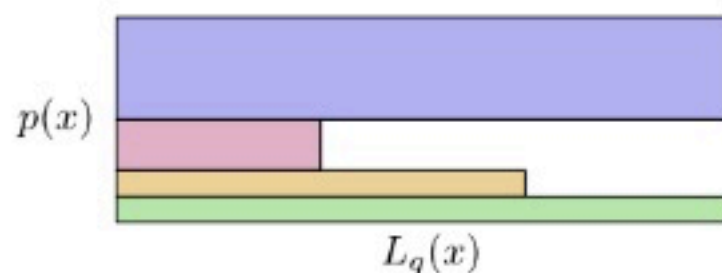
Average Length
of message from $q(x)$
using code for $p(x)$.

- Bob using his own code ($H(p) = 1.75$ bits)
- Alice using Bob's code ($H_p(q) = 2.25$ bits)
- Alice using her own code ($H(q) = 1.75$ bits)
- Bob using Alice's code ($H_q(p) = 2.375$ bits)

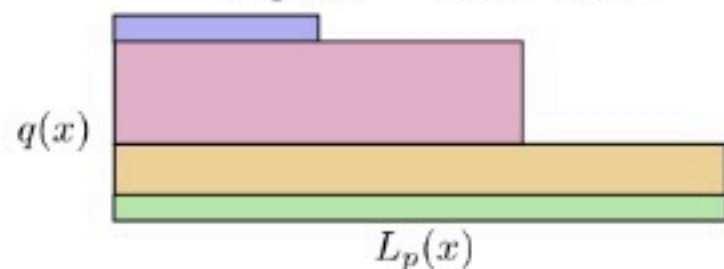
$$H(p) = H_p(p) = 1.75 \text{ bits}$$



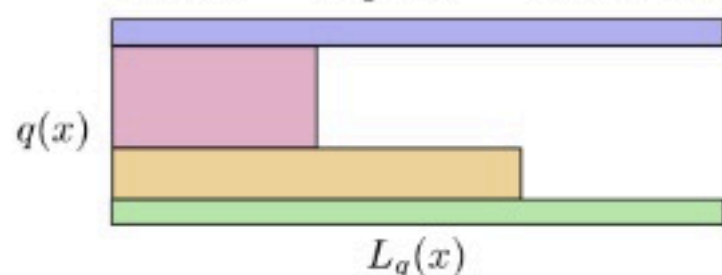
$$H_q(p) = 2.375 \text{ bits}$$



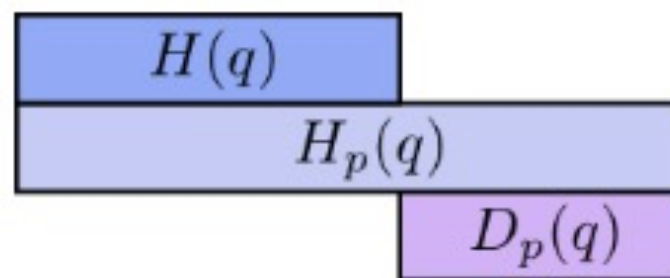
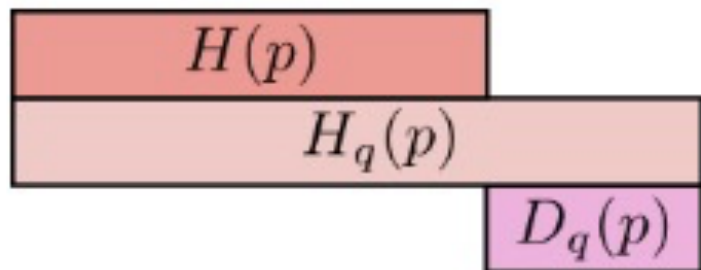
$$H_p(q) = 2.25 \text{ bits}$$



$$H(q) = H_q(q) = 1.75 \text{ bits}$$



相対エントロピー $D_q(p)$ (Kullback–Leibler divergence)



$$D_q(p) = H_q(p) - H(p)$$

相対エントロピー

条件付きエントロピー

□ Def. 1:

通信チャンネルは、入力状態 $X = \{x_1, \dots, x_\ell\}$ を出力状態 $Y = \{y_1, \dots, y_\ell\}$ にマップするデバイスである。

□ Def. 2:

X, Y 上の確率分布 $p(x_i)$ and $p(y_i)$ と結合分布 $p(x_i \wedge y_i)$ があたえられた時、条件付き確率 $p(x_i | y_j)$ は、次の式で定義される。

$$p(x_i | y_j) = p(x_i \wedge y_j) / p(y_j)$$

□ Def. 3:

Y が与えられた時、 X の条件付きシャノンエントロピーは、次の式であたえられる。

$$H(X|Y) = \sum_j p(y_j) \left(- \sum_i p(x_i | y_j) \log_2 p(x_i | y_j) \right)$$

$$H(X|Y) = \sum_j p(y_j) \left(- \sum_i p(x_i|y_j) \log_2 p(x_i|y_j) \right)$$

$$H(X|Y) = \sum_y p(y) \sum_x p(x|y) \log_2 \left(\frac{1}{p(x|y)} \right)$$

$$= \sum_{x,y} p(x,y) \log_2 \left(\frac{1}{p(x|y)} \right)$$

条件付きエントロピーの意味

- $H(X|Y)$ は、ある出力 Y の値が与えられた時、ある入力 X の値の平均の不確実さをあたえる。
- 我々が、このチャンネルを、非常に大きな回数 N だけ使ったとしよう。
- この時、典型的な入力文字列は、 $2^{NH(X)}$
- また、典型的な出力文字列は、 $2^{NH(Y)}$
- $X \wedge Y$ の、典型的な出力文字列は、 $2^{NH(X \wedge Y)}$

$$\left(\begin{array}{l} \text{出力}Y\text{の結果を出しうる} \\ \text{典型的な入力文字列の} \\ \text{数} \end{array} \right) = \frac{2^{NH(X \wedge Y)}}{2^{NH(Y)}} = 2^{N(H(X \wedge Y) - H(Y))}$$

条件付きエントロピーの意味

$$\textit{Claim: } H(X \wedge Y) = H(Y) + H(X|Y)$$

$$\textit{So: } \left(\begin{array}{c} \text{出力Yの結果を出しうる} \\ \text{典型的な入力文字列の} \\ \text{数} \end{array} \right) = 2^{NH(X|Y)}$$

- もし、ノイズのあるチャンネルでメッセージを送ろうとするなら、条件付きエントロピーは、ノイズに基づくエラーを訂正するために、補助的なノイズのないチャンネルで送られる必要のある、1文字あたりのbits の数を決める。

Check Claim:

$$H(X \wedge Y) = - \sum_{i,j} p(x_i \wedge y_j) \log_2[p(x_i \wedge y_j)]$$

$$= - \sum_{i,j} p(x_i \wedge y_j) \log_2[p(y_j)p(x_i|y_j)]$$

$$= - \sum_{i,j} p(x_i \wedge y_j) \log_2[p(y_j)] - \sum_{i,j} p(x_i \wedge y_j) \log_2[p(x_i|y_j)]$$

$$= - \sum_j p(y_j) \log_2[p(y_j)] - \sum_{i,j} p(y_j)p(x_i|y_j) \log_2[p(x_i|y_j)]$$

$$= H(Y) + H(X|Y)$$

$$\sum_i p(x_i \wedge y_j) = p(y_j)$$



21世紀 エンタングルメントとエントロピー

ブラックホールのエントロピーの発見

“Black holes and Entropy”

Jacob D. Beckenstein

<http://www.phys.huji.ac.il/~bekenste/PRD7-2333-1973.pdf>

1973年

ブラックホール「No Hair 定理」

- ブラックホールは、観測可能な物理量としては、質量・電荷・角運動量の三つしか持たないと信じられてきた。ブラックホールは、極めて単純な構造の存在で、上の三つの物理量でその特徴を、完全に記述できる。それを、物理学者のWheelerは、「ブラックホールには毛がない」といった。
- ブラックホールには、マイクロステーツがない？
- ベッケンシュタインの発見は、これを覆すもの。彼は、**ブラックホールのエントロピー S_{BL} が、その地平の表面積 A に比例することを発見する**。比例定数は、のちにホーキングが1/4であることを見つける。

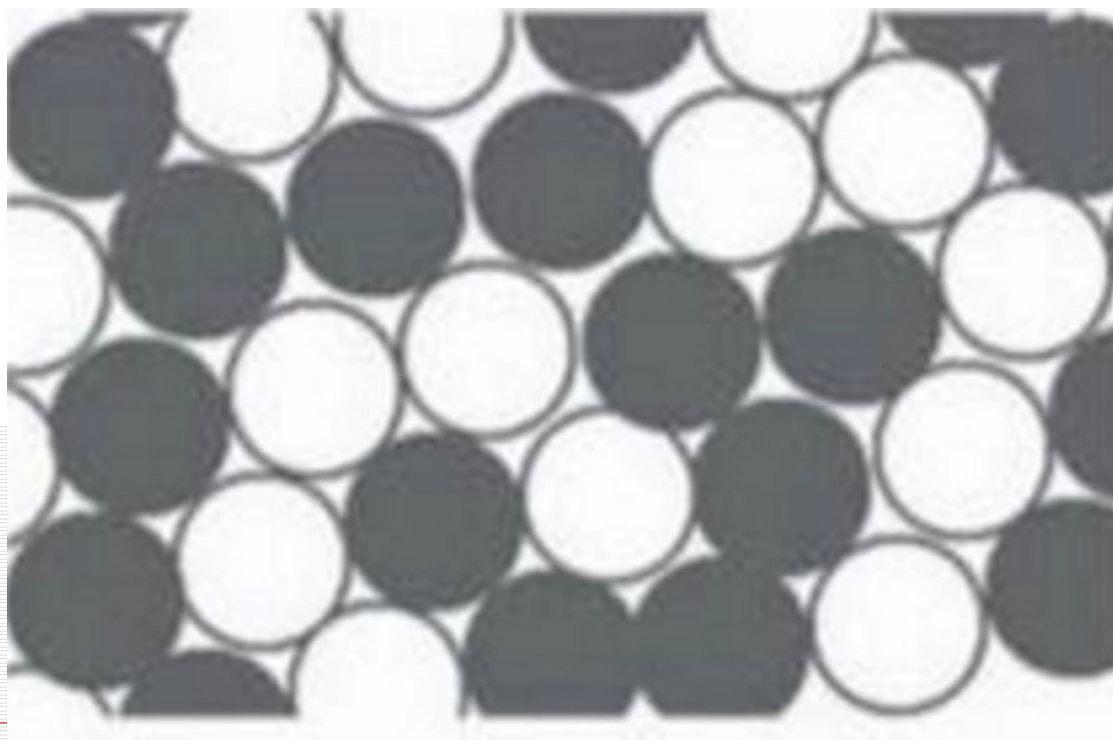
$$S_{BL} = \frac{1}{4} A$$

サスキンドの計算

- 今、ブラックホールに、光子(光の粒子)が一個落ち込んだとする。この光子は、1bitの情報を持っているとする。
(本当は、光子一個は、無数の情報を持ちうるのだが。光子の波長を、ブラックホールの地平の半径と同じ程度に広いものにすれば、光子の輪郭はぼやけ、それが持つ情報を小さなものにできる)
- この時、ブラックホールのエネルギーは、どれほど増えるだろうか？ それは、ブラックホールに落ち込んだ、1bitの情報を運ぶ光子のエネルギーだけ増える。その量は計算できる。
- エネルギーは質量に等しいので、この時、ブラックホールの質量が、どれだけ増えるかがわかる。
- ブラックホールの地平の半径 R は、ブラックホールの質量 M で決まる。半径がわかれば、その地平の面積の変化を計算できる。

- 1bitの情報を持つフォトン一個が、ブラックホールに落ち込むと、ブラックホールの質量は、 10^{-45} キログラム増えるという。小数点のゼロの下にゼロが45個並ぶ数だから、普通なら、無視しても構わない変化だ。
 - この時、地平の半径は、 10^{-72} メートル伸びるという。さっきの変化より、もっともっと小さい。普通なら、無視しても構わない変化だ。
 - この時、地平の面積は、 10^{-70} メートル広くなる。これも、無視して構わない変化だ。
 - サスキンドは、ここであることに気づく。 10^{-70} というのは、プランクの長さ $h = 10^{-35}$ の二乗じゃないか？
 - そこで、ブラックホールの質量を、いろいろ変えて計算すると、驚くべきことに、どんな質量のブラックホールでも、1bitの情報を持つフォトンが飛び込むと、その地平の面積は h^2 プランク長の二乗分だけ広がるのである！（一部、脚色あり。）
-

- イメージとしては、こういうこと。
- ブラックホールの地平が、 $h \times h$ のマス目で、びっしり埋められているとする。ここに、1bitの情報をもった光子が、飛び込むと、新しいマス目が、他を押しつけて一つ増えるということ。なんか、格好いい発見である。



新しい情報 / エネルギー観

- 落ち込んだフォトンの情報は、地平の平面にへばりつくのだ。
 - これは、「ブラックホールのエントロピーは、地平の面積に比例する。」という、ベッケンシュタインの重要な発見を説明し、さらに詳しく言い換えたものだ。
 - 「ブラックホールのエントロピーは、bitで計測すれば、プランクの面積単位で計測した地平の面積に比例する。」
 - 「温度とは、1bitの情報が追加された時の、システムのエネルギーの上昇である。」
 - これは、19世紀のボルツマン、20世紀のシャノンを超える、おそらく、21世紀の科学を特徴付ける、新しい 情報 / エネルギー観である。
-

ブラックホール輻射

S. W. Hawking

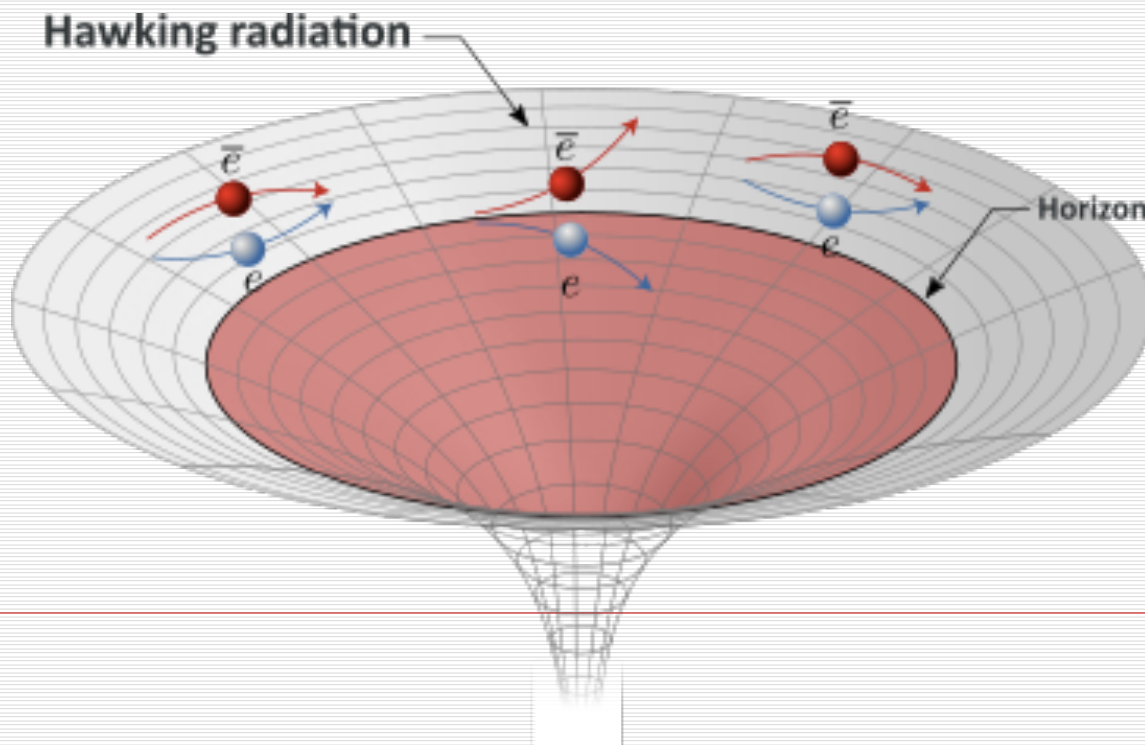
“Black hole explosion?”

<https://www.nature.com/nature/journal/v248/n5443/abs/248030a0.html>

1974年

ブラックホール輻射とブラックホールの蒸発

- ブラックホールの地平周辺で対発生した粒子の一方がブラックホールに落ち込み、た方がそこから抜け出すと、ブラックホールから粒子が飛び出すように見える。ブラックホールは、それによってエネルギーを失って、じょじょに蒸発していく。



ブラックホールは、「温度」も持つ

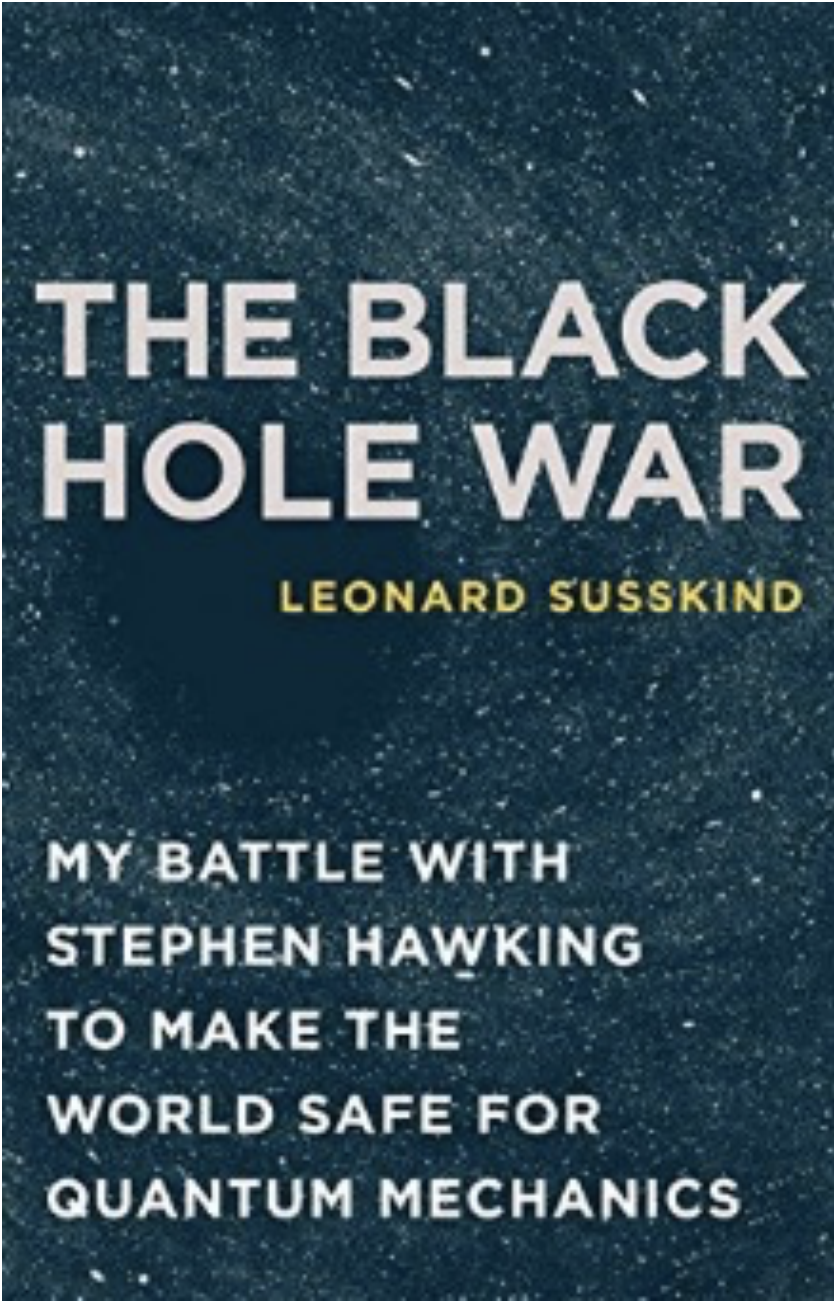
- この複写により、ブラックホールは、温度も持つことになる。ホーキングは、その温度も導出した。ブラックホールは質量 M が小さければ小さいほど高温であるといえる。とはいえその温度は、例えば太陽の数倍の質量を持つブラックホールの場合、100万分の1 K 程度となり、通常の恒星質量クラスのブラックホールでは宇宙背景放射の温度 (3 K) よりもずっと低い。



A diagram showing a central circle representing a black hole. Several wavy lines radiate outwards from the circle, representing the emission of Hawking radiation.

$$T = \frac{hc^3}{16\pi^2 G M k}$$

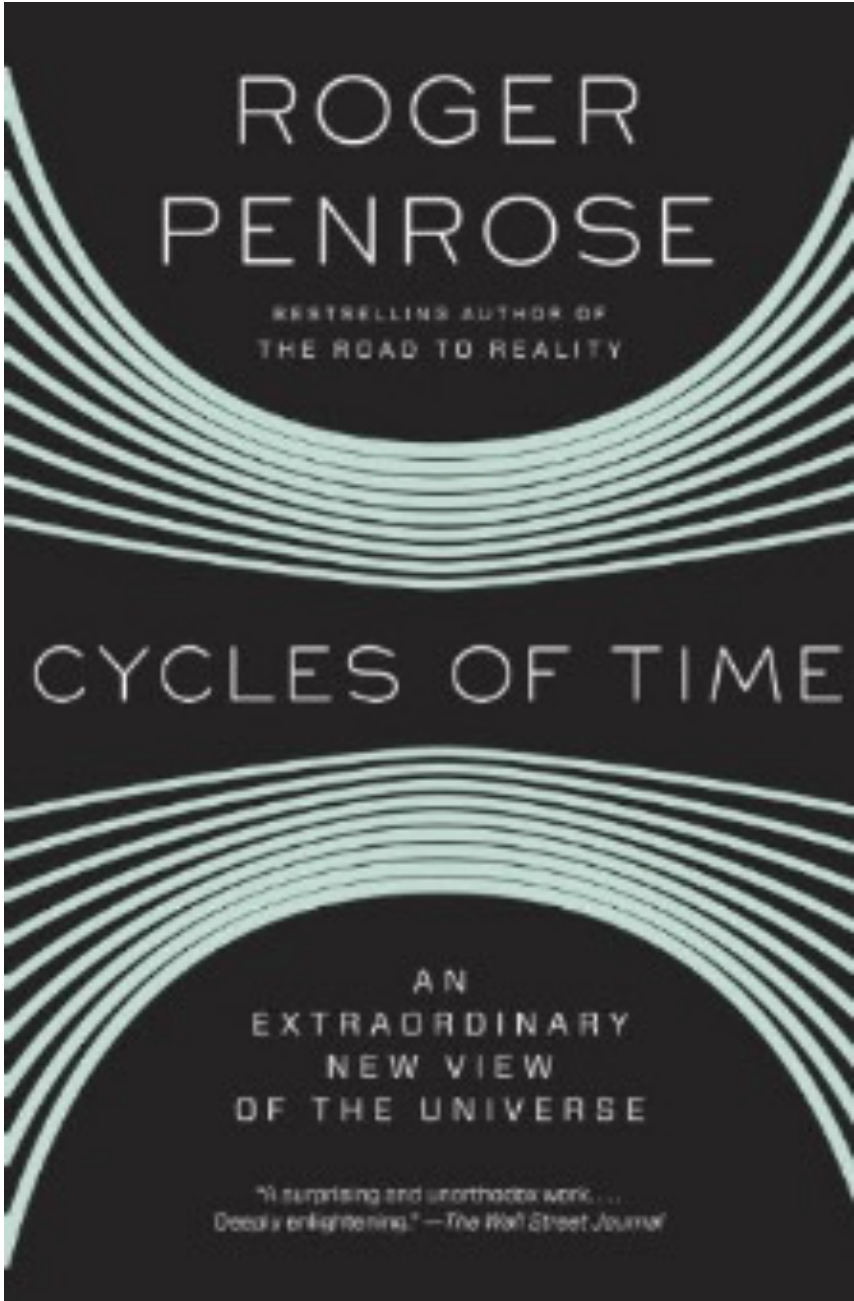
ブラックホールで情報は消滅するか



THE BLACK HOLE WAR

LEONARD SUSSKIND

MY BATTLE WITH
STEPHEN HAWKING
TO MAKE THE
WORLD SAFE FOR
QUANTUM MECHANICS



ROGER
PENROSE

BESTSELLING AUTHOR OF
THE ROAD TO REALITY

CYCLES OF TIME

AN
EXTRAORDINARY
NEW VIEW
OF THE UNIVERSE

"A surprising and unorthodox work...
Deeply enlightening." —The Wall Street Journal

AdS/CFT 対応

J. M. Maldacena,

“The large N limit of superconformal field theories
and supergravity”

<https://arxiv.org/pdf/hep-th/9711200.pdf>

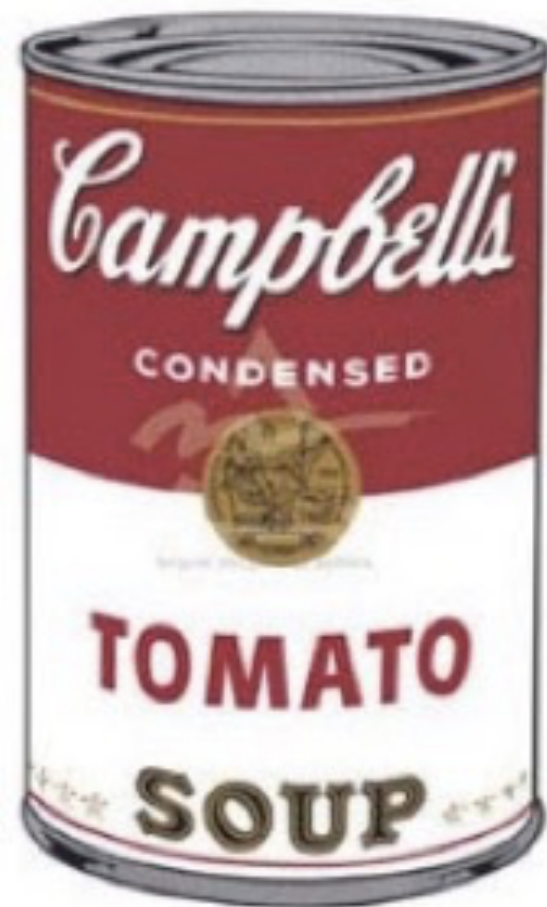
1998年1月

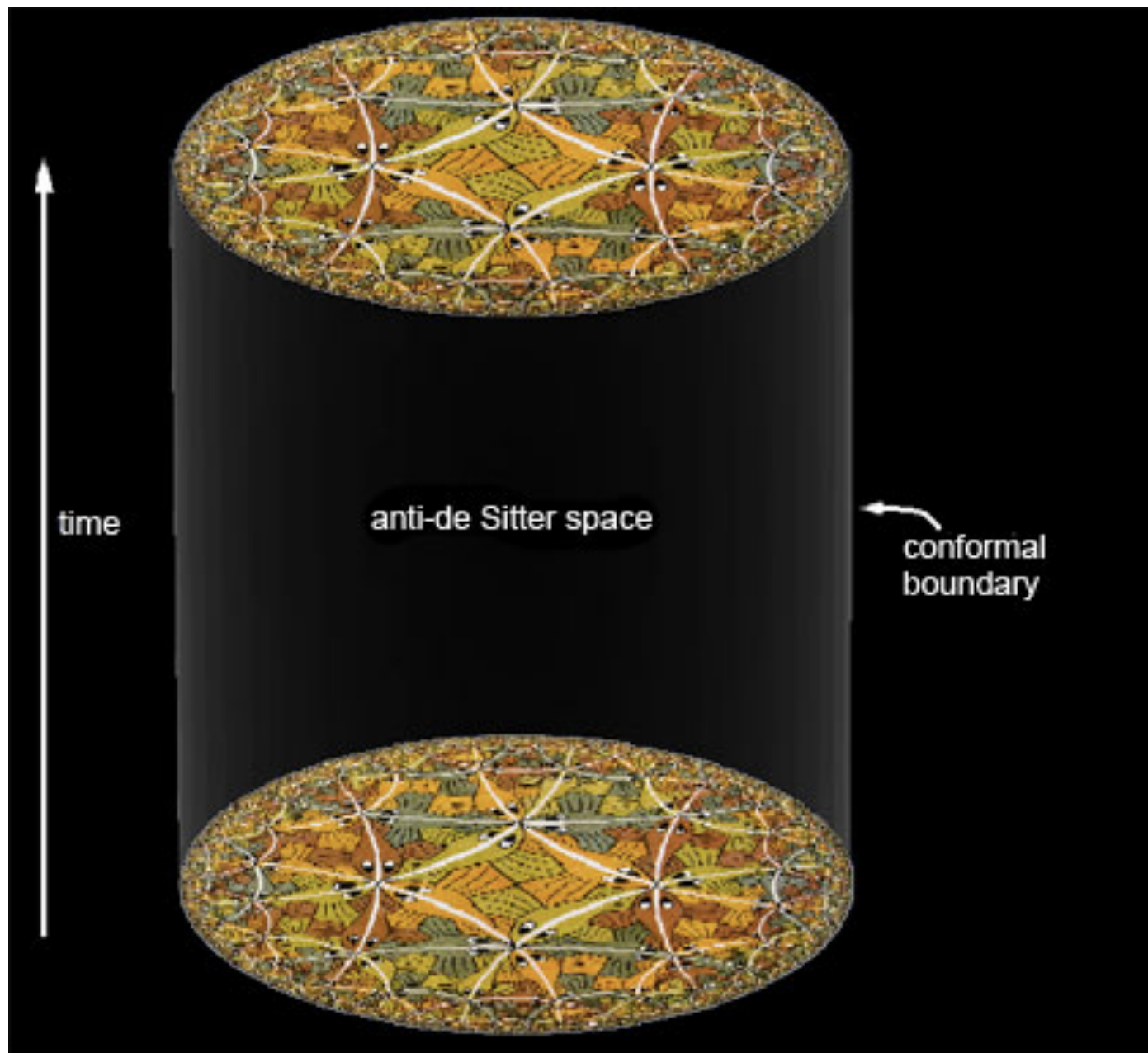
量子論と相対論の統一の課題

- 極微な世界を記述する量子力学と巨大な時空を記述する重力の理論である相対論との統一は、物理学の難問である。二つの理論は、それぞれの領域では、100年近くの間、大きな成功をおさめてきたのだが。それぞれ異なる理論体系に基づく、物理学が二つある？ それは、奇妙なことだ。
- このところ、この問題が解決できるのではという希望が生まれているように見える。もっとも重要な飛躍は、物理学者マルデセーナによる次のような発見である。(だいぶ、テキトーにまとめている)
- 「キャンベルのスープ缶にたとえば、重力理論はスープの理論で、量子論は缶の理論である。」
- ブリキの缶の外側をいくらなぞっても、スープのことはわからない。逆に、いくらスープを舐めても、缶のことはわからない。

量子論と相対論の思わぬ接点

- ❑ マルデセーナがエライのは、そうではないことを見つけたことである。「スープと缶は、無関係ではなく関係がある。缶を調べれば、スープのことがわかり、スープを調べれば缶のことがわかる！」
- ❑ それは、文字通り「接点」だった。重力理論(スープ)に「境界」を接して量子論(缶)が、棲んでいたのだから。これは、「境界」の重要性の発見でもあった。





<https://goo.gl/imCz3r>

エンタングルメントのエントロピー

Ryu and Takayanagi,

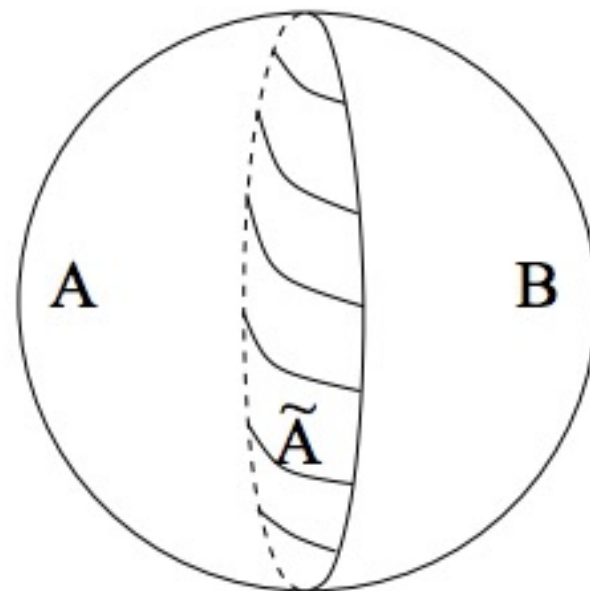
“Holographic derivation of entanglement entropy
from AdS/CFT”

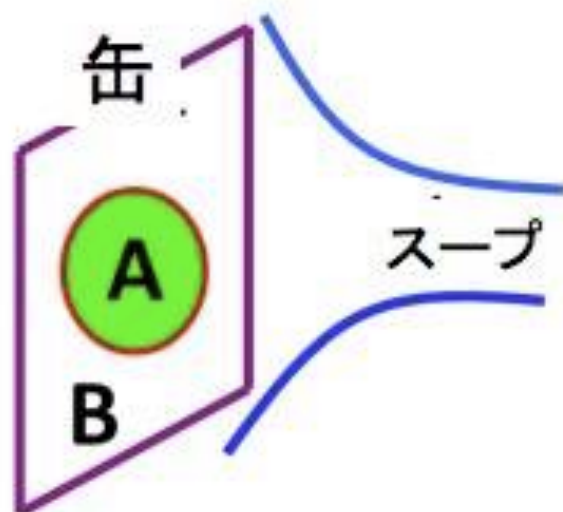
<https://arxiv.org/pdf/hep-th/0603001.pdf>

2006年5月

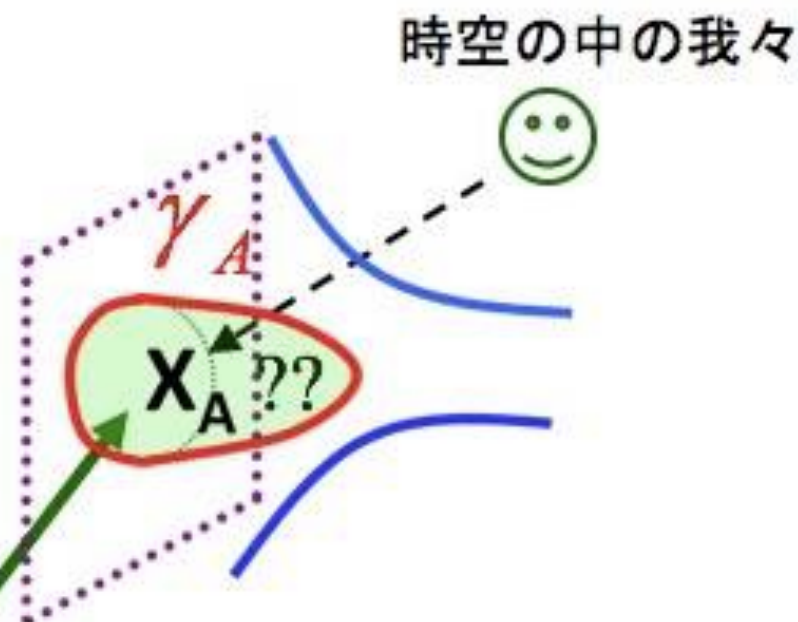
日本人物理学者 笠と高柳

- 次に、日本の二人の物理学者、笠と高柳が登場する。
- 「スープ(「時空」だと思っていい)が二つの部分 AとBに別れているとする。AとBの「境界部分」は、スープの「境界」だから、マルデセーナの理論にしたがって量子論で記述できるはず。」
- 「やってみたら、この「境界」は、なんと、量子論の不思議な「もつれあい(エンタングルメント)」のエントロピーに対応するんだ！」
- 彼らは、「量子もつれあい」がエントロピーを持つことの、最初の発見者だ。これは、ベッケンシュタインによる、ブラックホールがエントロピーを持つことの発見に比肩できる発見だ。





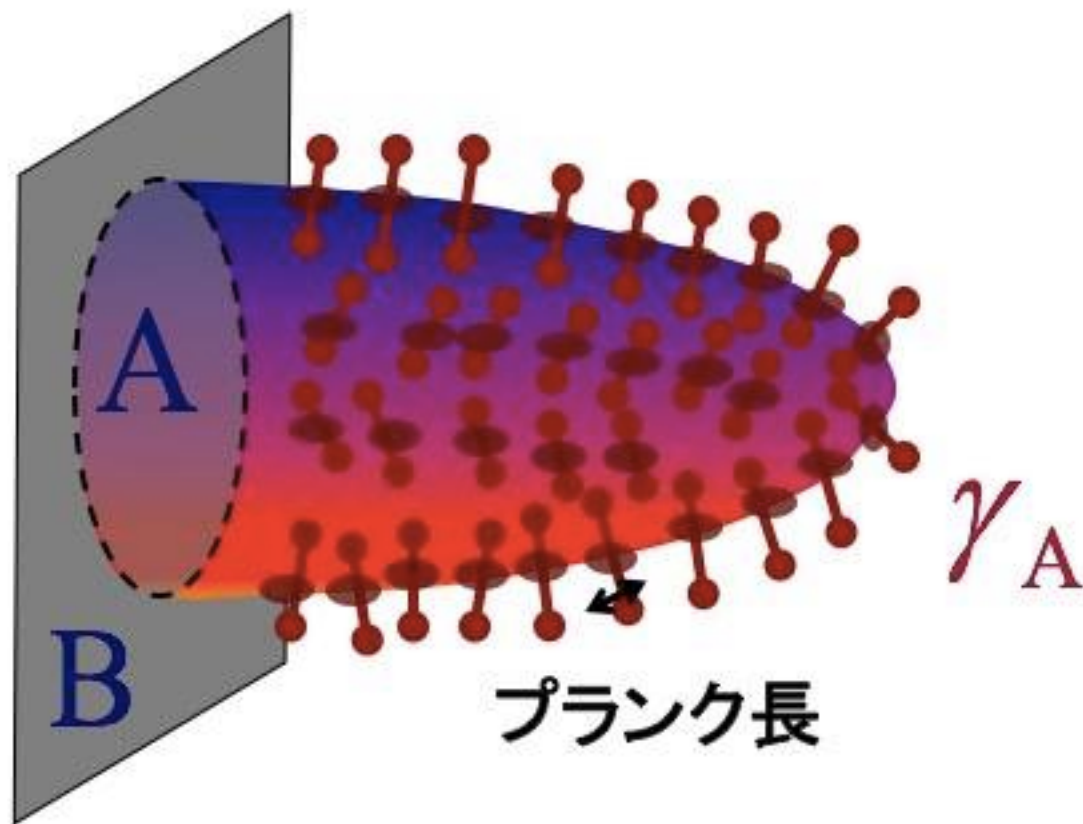
Aは、缶上の領域



時空中の我々

X_A は、スープ上の領域

Aの情報は、ここに
エンコードされる



$$S_A = \frac{\text{Area}(\gamma_A)}{4G_N}$$

$$\approx \frac{\text{Area}(\gamma_A)}{l_{pl}^2}.$$

時空を結びつけるエンタングルメント

M. Van Raamsdonk,

“Building up spacetime with quantum entanglement”

<https://arxiv.org/pdf/1005.3035.pdf>

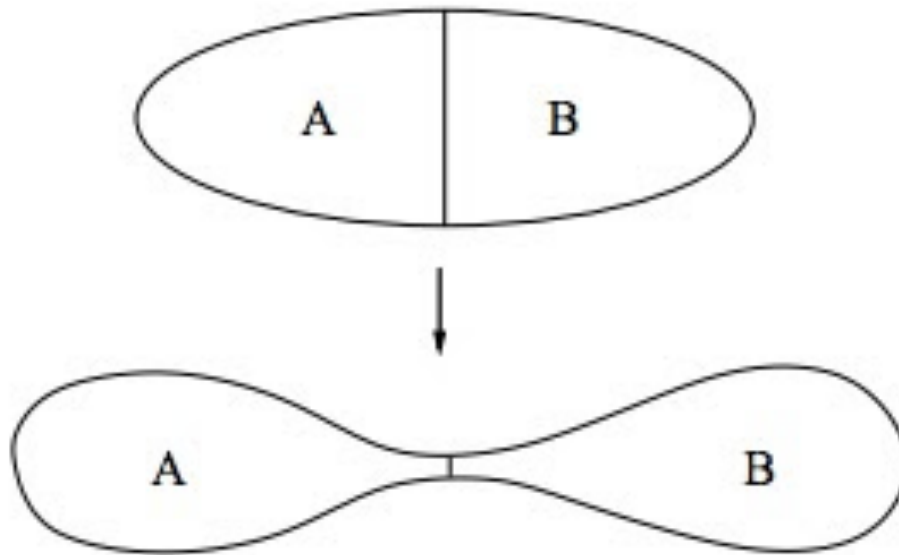
2010年5月

エンタングルメントのエントロピーが 時空を縫い合わせる！

- これに刺激されて、今度は、カナダのラームズダンクがくる。
(以下も、テキトー。もっとも、彼の最初の頃の論文は、テキトーなこと言っているとみなされて、どの学会誌からも、掲載を拒否されたらしい。)
 - 「二つの時空 A , B があったとする。二つの時空を、引き離してみる。そうすると、「境界面」の面積は減少する。これは、二つの時空が離れれば離れるほど、もつれあいのエントロピーが減ることを意味している。このエントロピーがゼロになった時、二つの時空は、引きちぎられる。」
-

エンタングルメントのエントロピーが 時空を縫い合わせる！

- 「そうだ！ 逆に考えればいいんだ。時空を結びつけているのは、もつれあいなんだ。エンタングルメントのエントロピーが、時空を縫い合わせているんだ。」
- これは、巨大な飛躍だった。





Tensor Network

Roman Orus

"A Practical Introduction to Tensor Networks:
Matrix Product States and Projected Entangled
Pair States"

<https://arxiv.org/pdf/1306.2164.pdf>

凝縮系物性論のチャレンジ

- 液体ヘリウムが、超低温状態で超電導・超流動の性質を示すことは、バーディーン、クーパー、シュリーファアのBCS理論で、電子等のフェルミ粒子が、凝縮してボース粒子の性質を示す（一つの系で、複数の粒子が同一の状態を示す）ことで説明される。凝縮系物性論は、こうした物質の新しい相を研究する。
 - 現在でも、超電導技術は、リニア新幹線や医療でのMRIのように広く利用され、また、量子コンピュータでも基本的な要素技術として利用される。凝縮系物性論は、テクノロジーを大きく変える可能性を持つ、チャレンジングな分野である。
 - ただ、現在では、高温超電導のような現象を、理論的に説明することはできない。
-

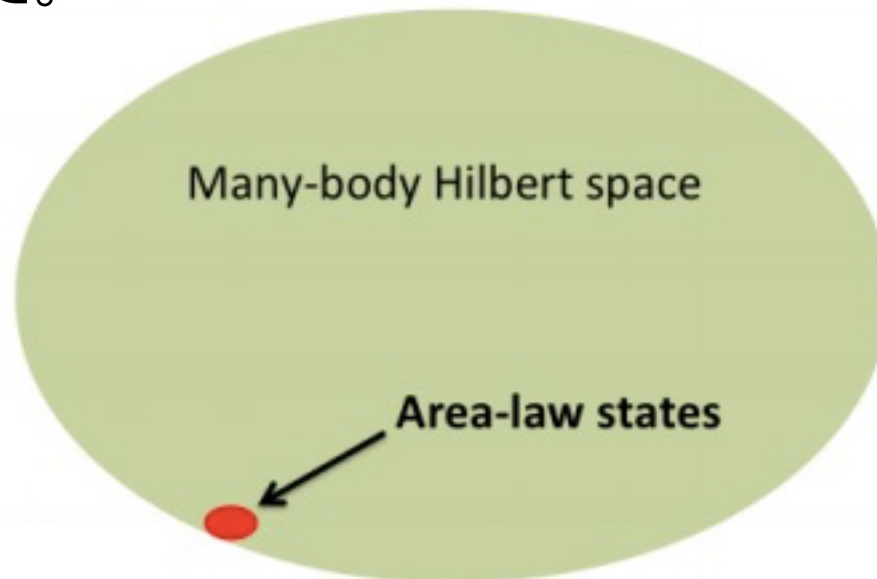
量子多体系の問題を解く難しさ

- 凝縮系物性論のターゲットとする領域を、「量子多体系」と呼ぶことがある。多体系の量子論的状态を明らかにするためには、粒子の数 N に対応して、(状態の数) N の次のような複雑な方程式をヒルベルト空間上で解く必要がある。
- $N=10^{23}$ だとしても、その複雑さは、 $10^{(10^{23})}$ にもなる。宇宙の素粒子の数が、 10^{90} 程度だというから、これを解くのは事実上、不可能である。

$$\begin{aligned} \langle \Psi | \hat{H} | \Psi \rangle = & \sum_{\sigma_1, \dots, \sigma_L} \sum_{\sigma'_1, \dots, \sigma'_L} \sum_{a_1, \dots, a_L} \sum_{a'_1, \dots, a'_L} \sum_{b_1, \dots, b_L} \\ & A_{1,a_1}^{*\sigma_1} W_{1,b_1}^{\sigma_1, \sigma'_1} A_{1,a'_1}^{\sigma'_1} A_{a_1, a_2}^{*\sigma_2} W_{b_1, b_2}^{\sigma_2, \sigma'_2} A_{a'_1, a'_2}^{\sigma'_2} \dots \\ & \dots A_{a_{L-2}, a_{L-1}}^{*\sigma_{L-1}} W_{b_{L-2}, b_{L-1}}^{\sigma_{L-1}, \sigma'_{L-1}} A_{a'_{L-2}, a'_{L-1}}^{\sigma'_{L-1}} A_{a_{L-1}, 1}^{*\sigma_L} W_{b_{L-1}, 1}^{\sigma_L, \sigma'_L} A_{a'_{L-1}, 1}^{\sigma'_L} \end{aligned}$$

エントロピーが「Area則」を満たす 基底状態の計算の簡素化が可能に

- 「Area則（面積則）」というのは、エントロピーが系の体積ではなく系の境界の面積に比例するという法則のこと。エンタングルメントのエントロピーは、Area則を満たす。
- この時、量子多体系の基底状態の解は、一般の巨大なヒルベルト空間の中ではなく、ごく限られた領域の中にあることが、明らかになった。



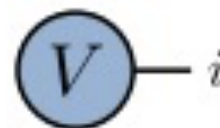
Tensor Networkが、グラフでその解を表現する

Notation

Tensor Network は、グラフとして表現される

ベクトル

$$V_i = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$



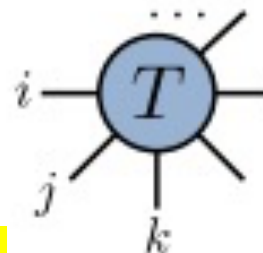
行列

$$M_{ij} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$



テンソル

$$T_{ijk\dots}$$



<https://goo.gl/Zq9WUw>

テンソルの縮約

つなげた線が、縮約を表す

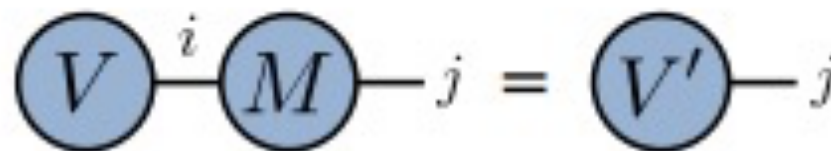
二つのベクトル

$$\sum_i V_i V'_i = S$$



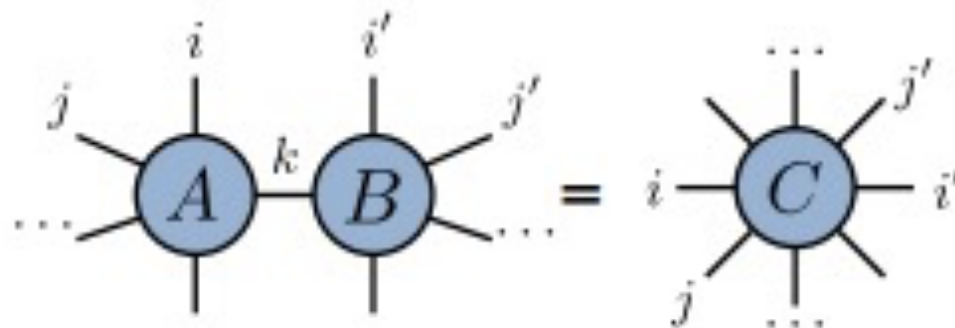
ベクトルと行列

$$\sum_i V_i M_{ij} = V'_j$$

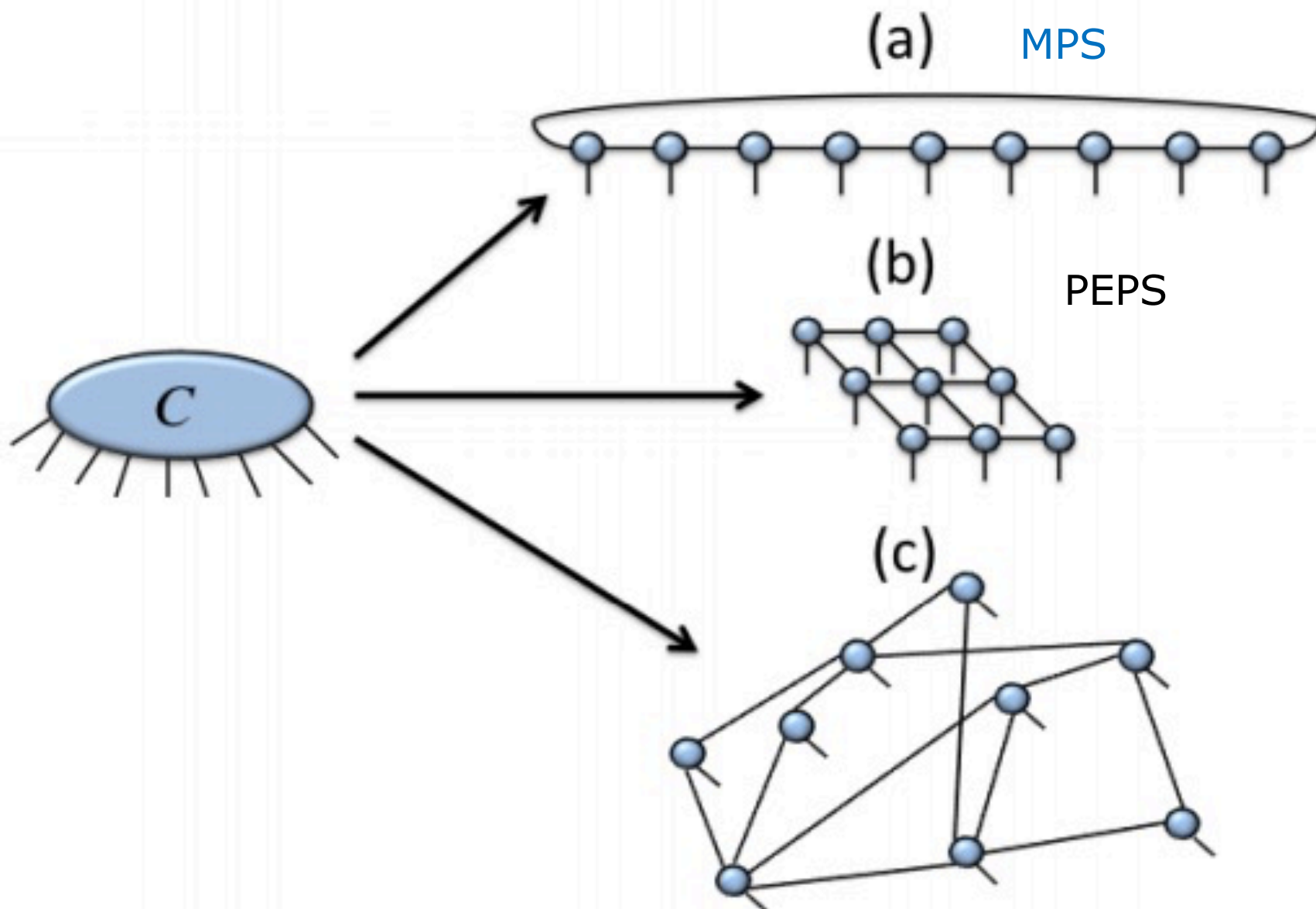


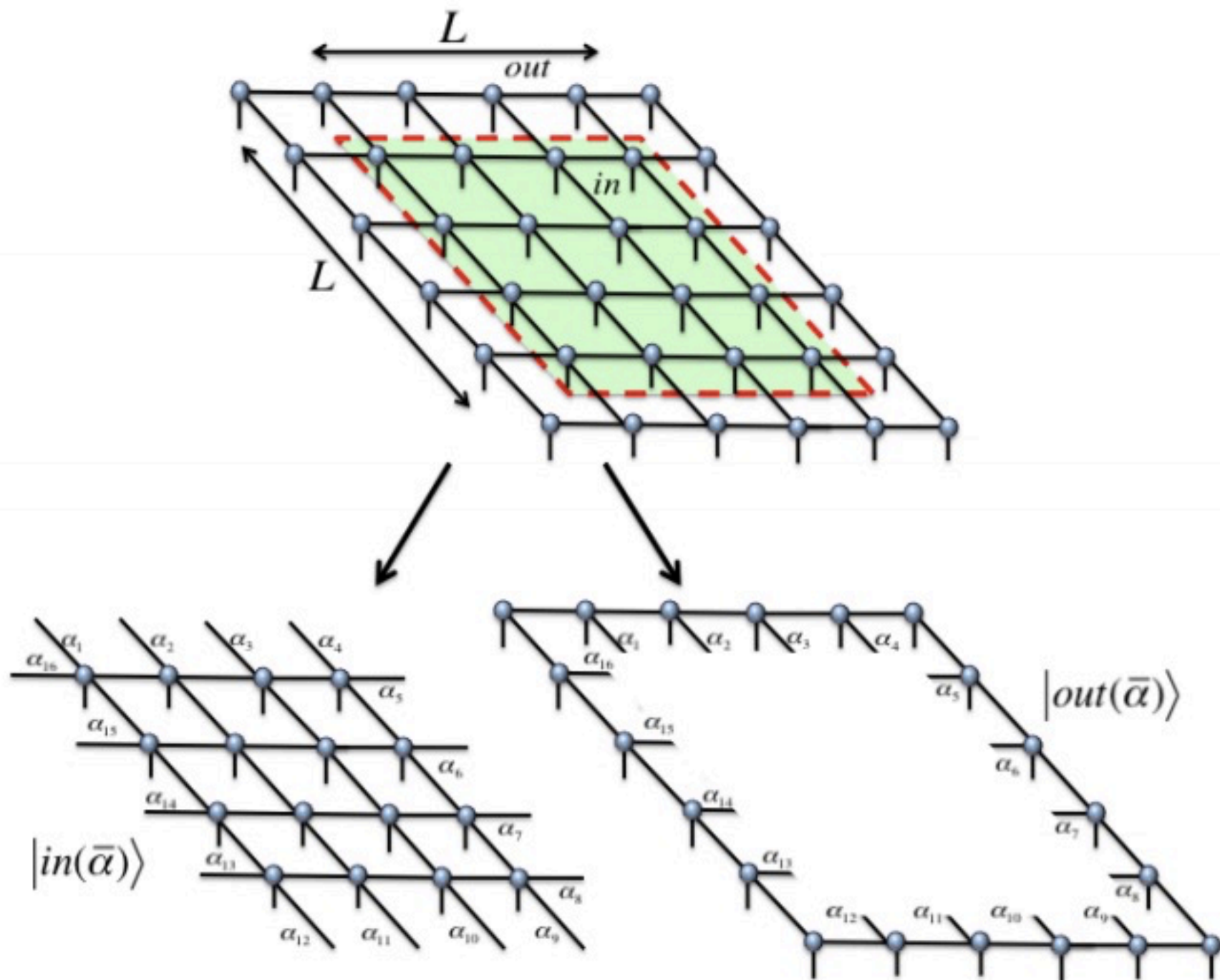
二つのテンソル

$$\sum_k A_{ij\dots k\dots} B_{i'j'\dots k\dots} = C_{ij\dots i'j'\dots}$$

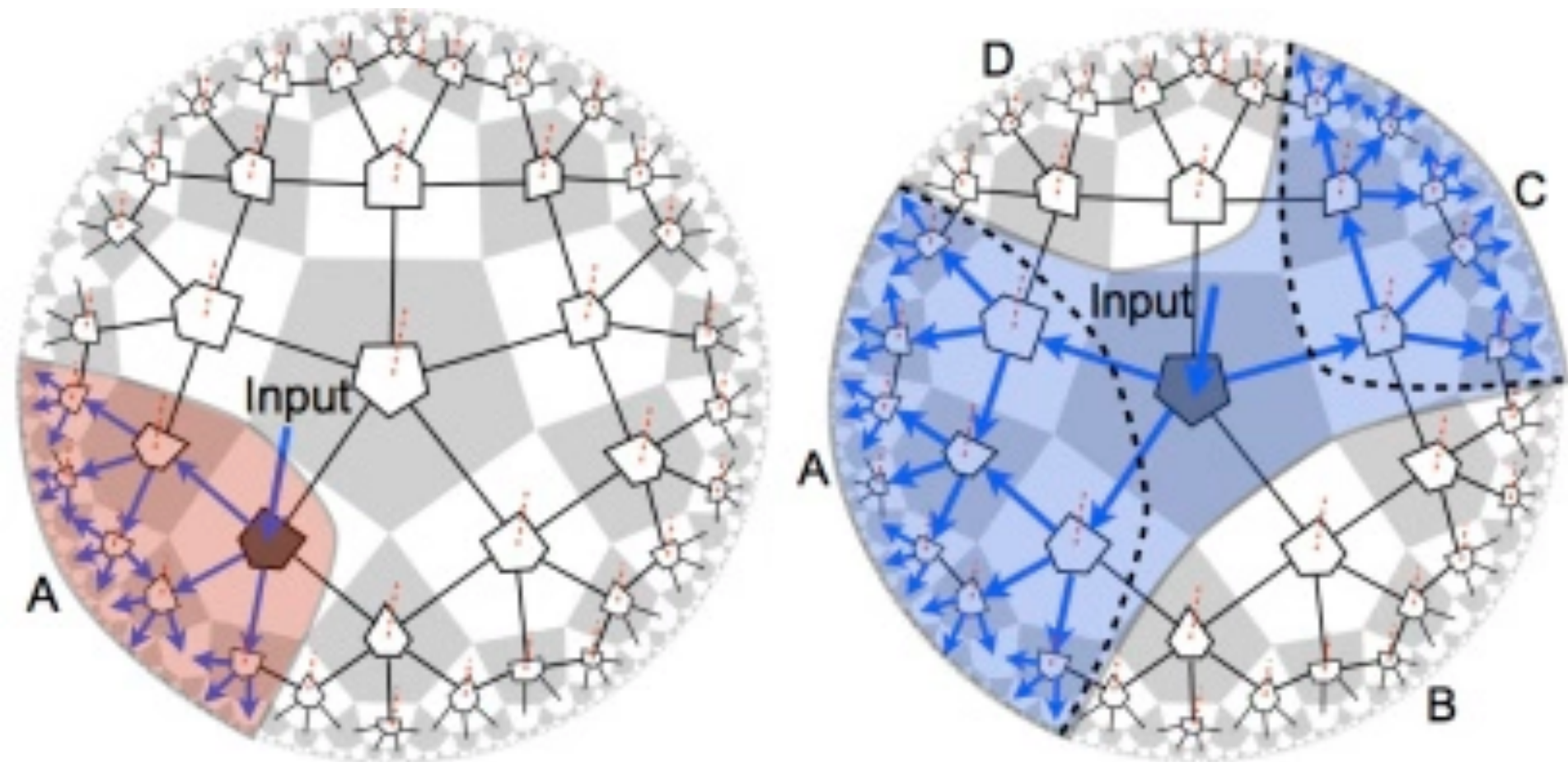


Tensor NetworkはテンソルCを、いくつかのパターンに分解する





AdS/CFT対応とTensor Network(MERA)



<https://goo.gl/GUCedR>

AMPS Paradox 2012

Scott Aaronson

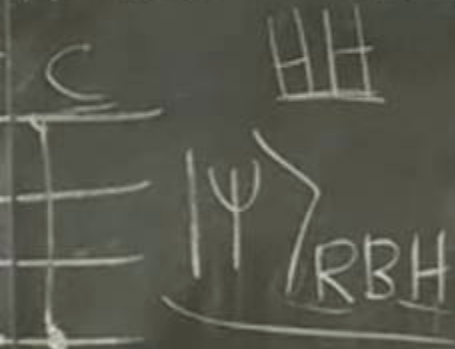
“The Cryptographic Hardness of Decoding Hawking Radiation”

<https://www.scottaaronson.com/talks/hawking.ppt>

2014年4月



HAYDEN ARGUMENT



Size (#gates) $\leq n^{O(1)}$

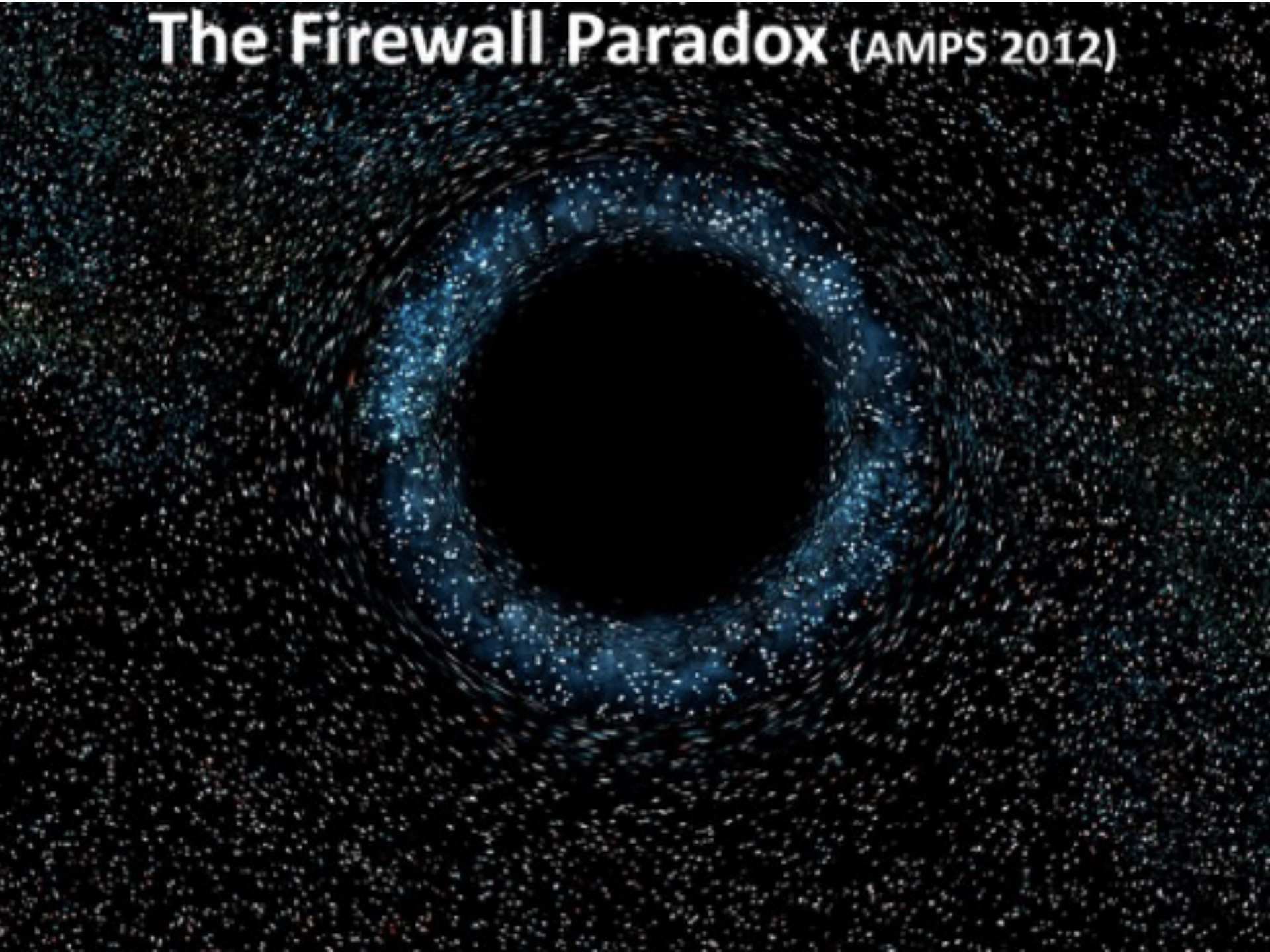
$$\rho_{BH} = I_{BH}$$

$\exists$ 1 qubit of
distillable entanglement
between R & B

Alice's task

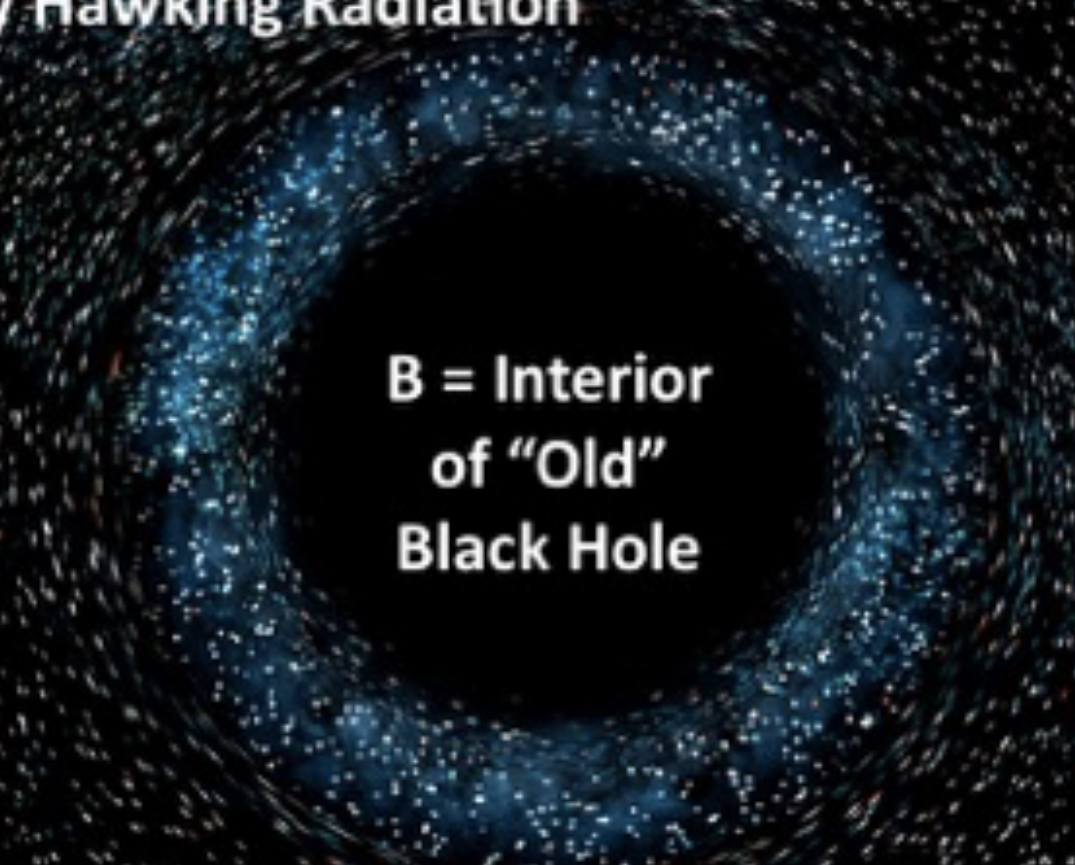
Apply U_R to R
so that

The Firewall Paradox (AMPS 2012)



The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation



B = Interior
of "Old"
Black Hole

The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation

H = Just-Emitted Hawking Radiation

B = Interior
of "Old"
Black Hole

A diagram illustrating the Firewall Paradox. It features a central black circle labeled 'B = Interior of "Old" Black Hole'. Surrounding this is a blue ring labeled 'H = Just-Emitted Hawking Radiation'. The entire structure is set against a background of a starry field, with a larger, fainter blue ring labeled 'R = Faraway Hawking Radiation' visible in the distance.

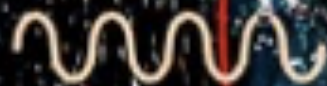
The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation

H = Just-Emitted Hawking Radiation

**B = Interior
of "Old"
Black Hole**

Near-maximal
entanglement



The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation

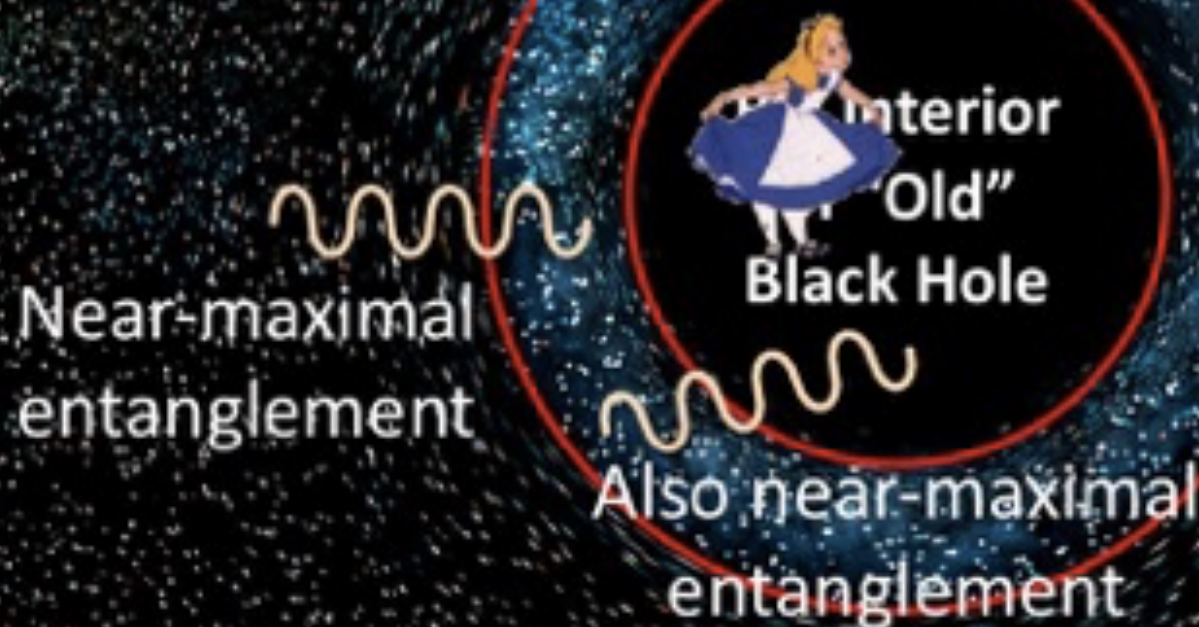
H = Just-Emitted Hawking Radiation



The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation

H = Just-Emitted Hawking Radiation



Violates monogamy of entanglement! The same qubit can't be maximally entangled with 2 things

The Firewall Paradox (AMPS 2012)

R = Faraway Hawking Radiation

H = Just-Emitted Hawking Radiation



Violates monogamy of entanglement! The same qubit can't be maximally entangled with 2 things

$$ER=EPR$$

Juan Maldacena & Leonard Susskind

"Cool horizons for entangled black holes"

<https://arxiv.org/pdf/1306.0533.pdf>

2013年7月

1935年のアインシュタインの二つの論文

EPR論文 エンタングルメントの発見

□ 「神はサイコロを振らない」

アインシュタインが、量子論に懐疑的だったことは、よく知られていると思う。量子論の否定は、現在の我々の目から見れば、明らかに、アインシュタインの「誤り」と言っているものだ。

□ 1935年に、アインシュタインとポドルスキーとローゼンは、概略、次の論文を発表する。(三人の著者の頭文字をとって、EPR論文と呼ばれる。) **彼が、次の論文で量子論の中に見つけた「矛盾」は、当時の物理学者の誰も気がつかなかった深いものだった。**

□ "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete ?" 「物理的な実在の量子力学の記述は、完全なものと考えることができるか？」 <https://goo.gl/qAWacP>

1935年のアインシュタインの二つの論文

EPR論文 エンタングルメントの発見

- アインシュタインが発見した、量子力学の「矛盾」を、「EPRパラドックス」という。これに「エンタングルメント」という名前をつけたのは、シュレディンガーらしい。
 - 今日では、エンタングルメントは、パラドックスどころか、量子論の基礎を構成するもっとも基本的な前提となっている。
 - ただ、このエンタングルメントという奇妙な現象を、誰よりも先に発見したのは、他ならぬアインシュタインであったことは、興味ふかいことである。
-

EPR論文を報ずる1935年のニューヨークタイムズ紙

EINSTEIN ATTACKS QUANTUM THEORY

Scientist and Two Colleagues
Find It Is Not 'Complete'
Even Though 'Correct.'

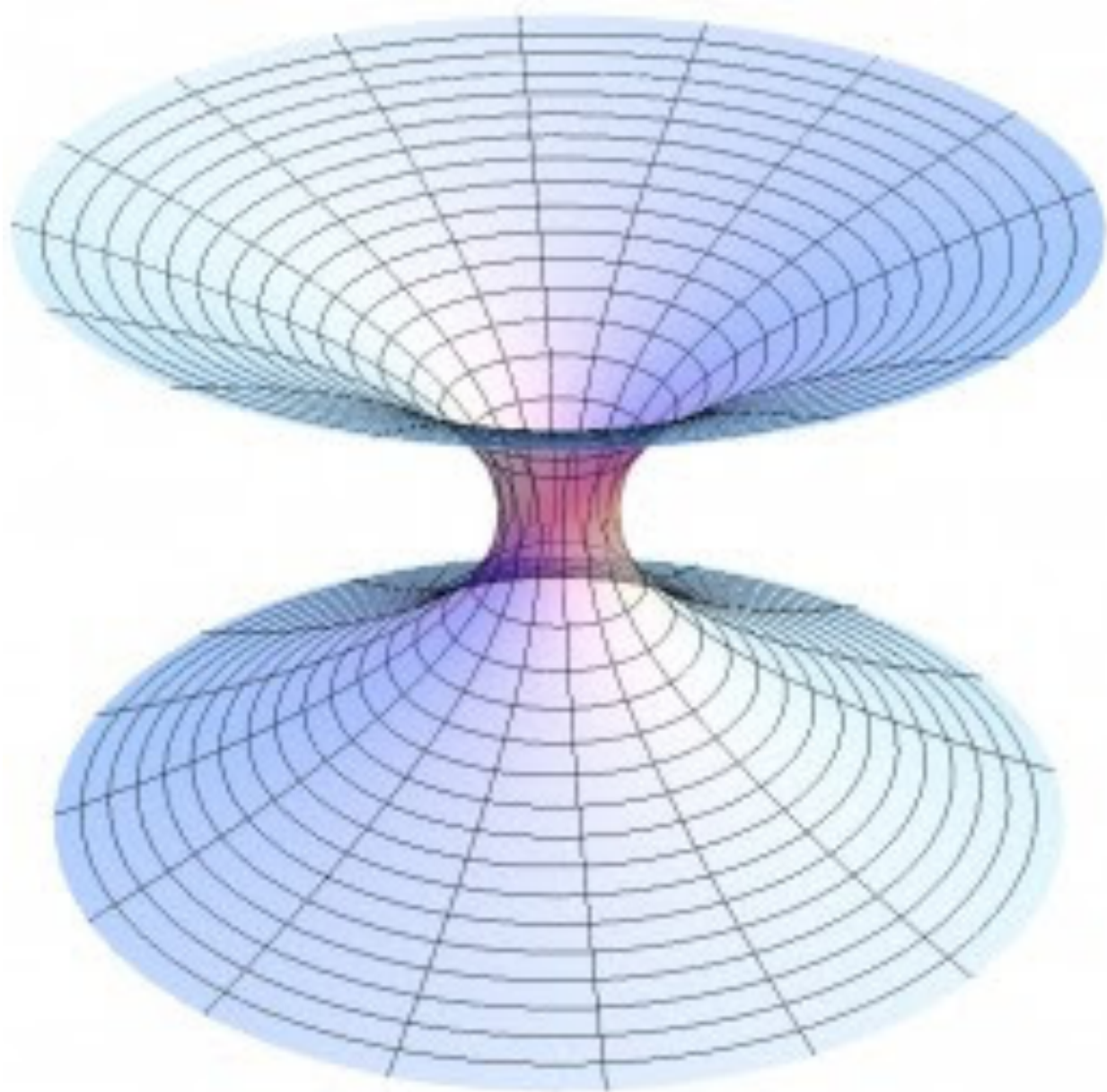
SEE FULLER ONE POSSIBLE

Believe a Whole Description of
'the Physical Reality' Can Be
Provided Eventually.

1935年のアインシュタインの二つの論文

ER論文 ブラックホールのワームホール

- アインシュタインは、先のEPR(Einstein, Podolsky, Rosen)論文を5月に公開してすぐに、ローゼンと一緒に、次の論文を公開する。1935年7月のことだ。この論文をER論文と呼ぶことがある。
 - "The Particle Problem in the General Theory of Relativity"「相対性の一般理論における粒子の問題」<https://goo.gl/KG4mjp>
 - 一般には、二つのブラックホールを結ぶ「橋」が存在しうることを見つけた論文だと言われている。この「橋」は、"Einstein-Rosen Bridge" と呼ばれる。別名「ワームホール」とも呼ばれる。(ただ、"wormhole" でWeb をググっても、あまりスジのいい情報は引っかからない。)
-



1935年のアインシュタインの二つの論文 ER論文とEPR論文には関連があるのか？

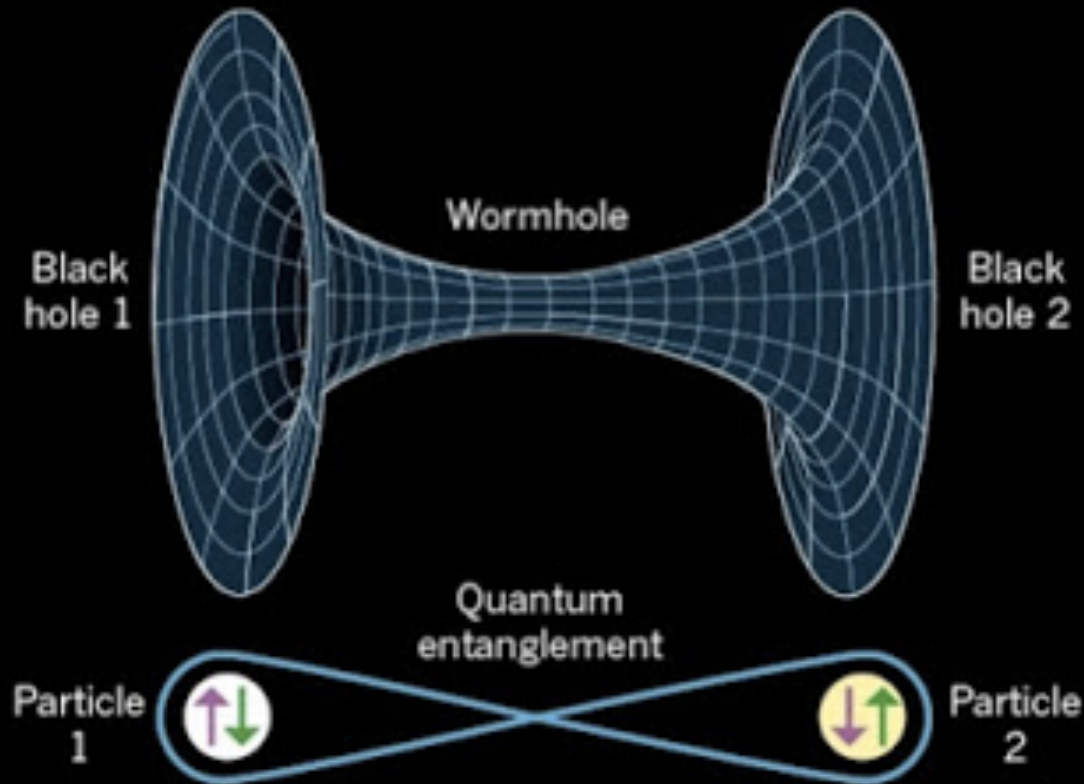
- 1935年に、アインシュタインは、「エンタングルメント」(ただし、パラドックスとして)と「ワームホール」を発見しているのである。ほぼ同時期に行われた、この二つの発見に何か関連があるのだろうか？
 - 80年前になされるべきこうした問いかけを、現代に蘇らせたのは、マルダセナ(Maldacena、AdS/CFT 対応の発見者)とサスキンド(Susskind)だった。2013年の論文 "Cool horizons for entangled black holes" <https://goo.gl/wU1pKK> で、二人は、大胆な仮説を提示する。
-

ER=EPR仮説

- 「二つのブラックホールの間のアインシュタイン・ローゼン・ブリッジは、二つのブラックホールのミクロな状態のEPR状の相関関係(エンタングルメント)によって生成される。... 我々は、これを $ER = EPR$ 相関関係と呼ぶ。別の言葉で言えば、ERブリッジは、その中では、量子システムに関連するEPRが、アインシュタインの重力の記述に弱く結合している特殊なタイプのEPR相関関係なのである。」
 - 要するに、1935年にアインシュタインが発見した「エンタングルメント」と二つのブラックホールを結ぶ「アインシュタイン・ローゼン・ブリッジ」は、スケールが全く違うのだが同じものだということである。これを、「 $ER = EPR$ 仮説」と呼ぶ。
-

ER = EPR

Also in 1935, Einstein and Rosen (ER) showed that widely separated black holes can be connected by a tunnel through space-time now often known as a wormhole.



Physicists suspect that the connection in a wormhole and the connection in quantum entanglement **are the same thing, just on a vastly different scale.** Aside from their size there is no fundamental difference.

Natureの2015年11月の記事 "The quantum source of space-time"
<https://goo.gl/x0HqA2>

Wormholes and Entanglement

John C. Baez & Jamie Vicary

Wormholes and Entanglement

<https://arxiv.org/pdf/1401.3416.pdf>

2014年2月

-
- Maldacena and Susskind have proposed a correspondence between wormholes and entanglement, dubbed ER=EPR. We study this in the context of 3d topological quantum field theory, where we show that the formation of a wormhole is the same process as creating a particle-antiparticle pair. A key feature of the ER=EPR proposal is that certain apparently entangled degrees of freedom turn out to be the same. We name this phenomenon 'fake entanglement', and show how it arises in our topological quantum field theory model.
-

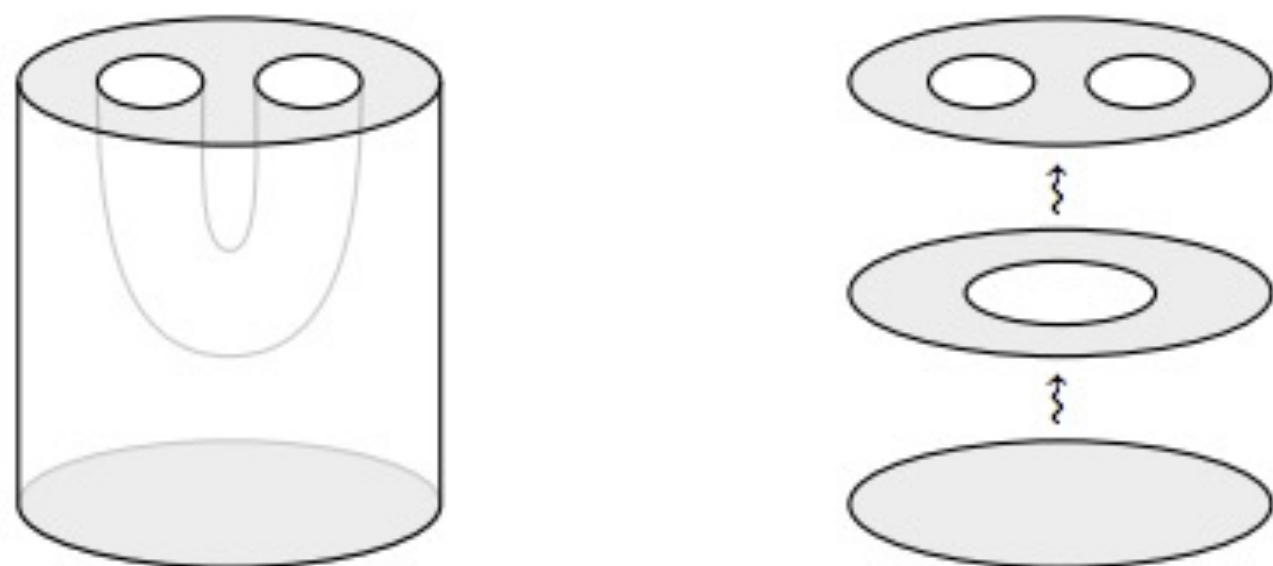


Figure 2: Spacetime history of pair creation in a 3d TQFT

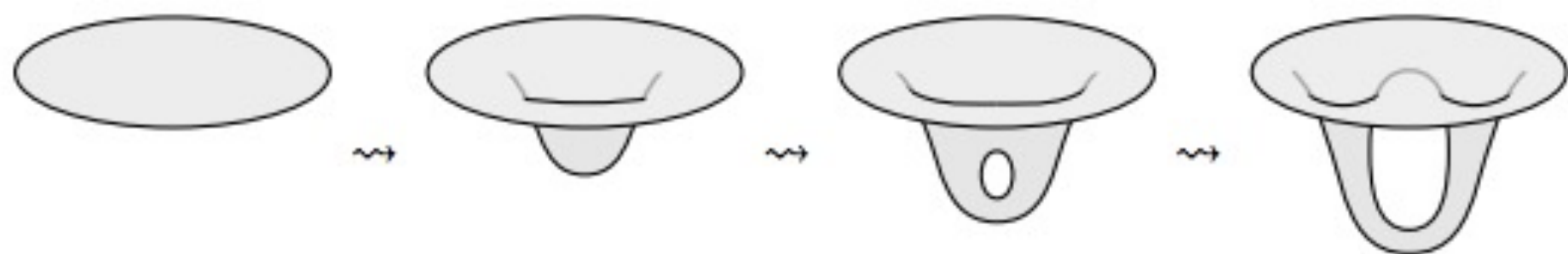
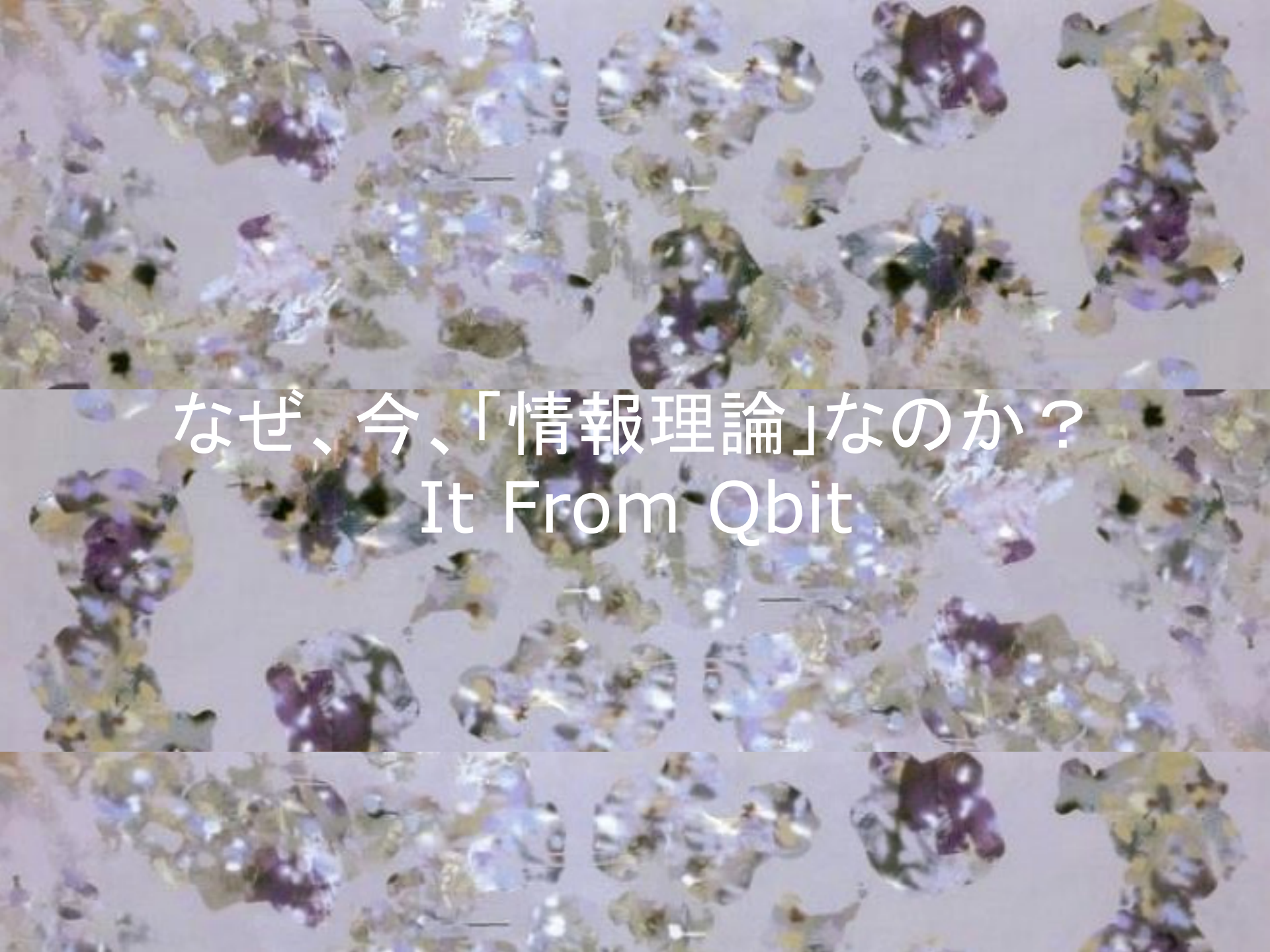


Figure 3: Formation of a wormhole

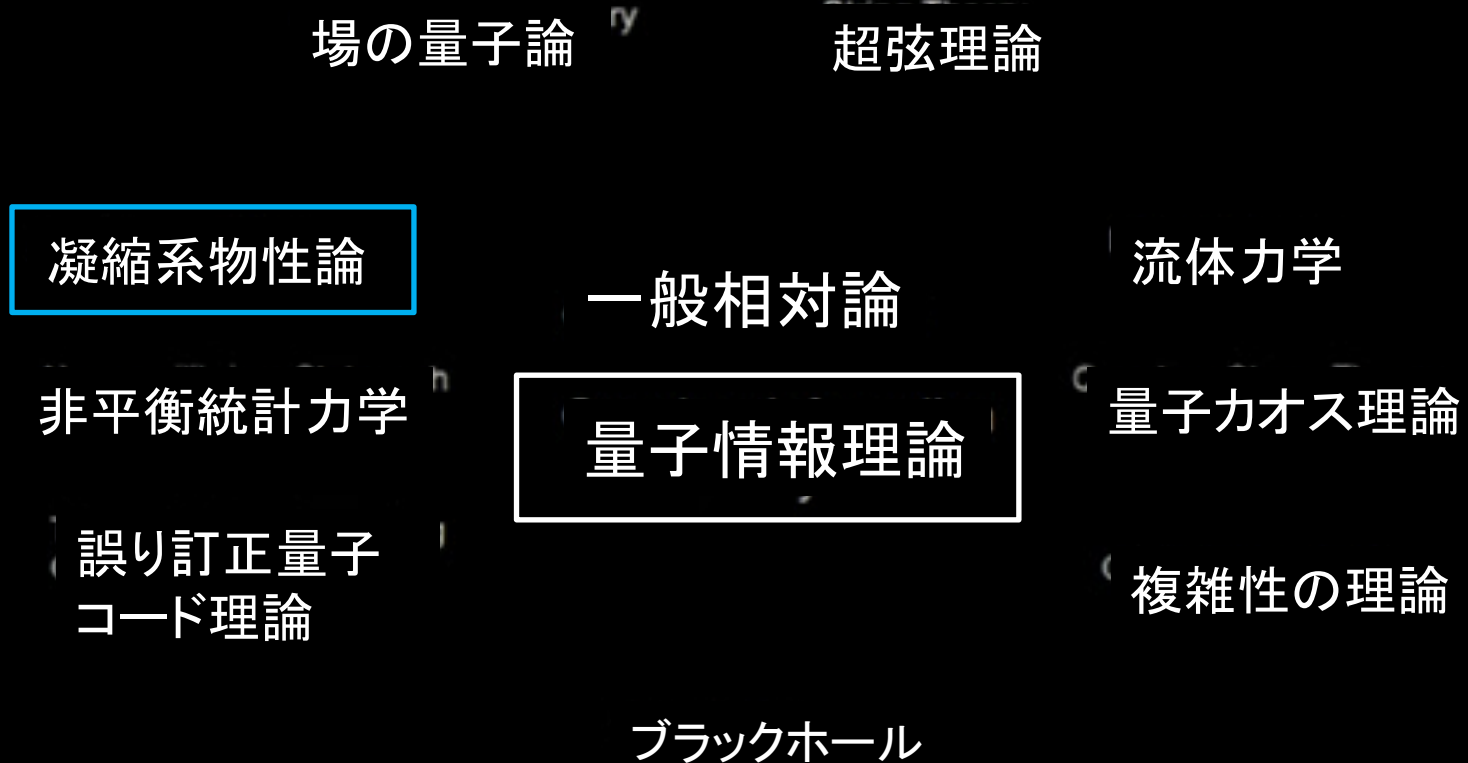


なぜ、今、「情報理論」なのか？
It From Qbit

物理学者は、 現代の物理学をどのように捉えているのか？

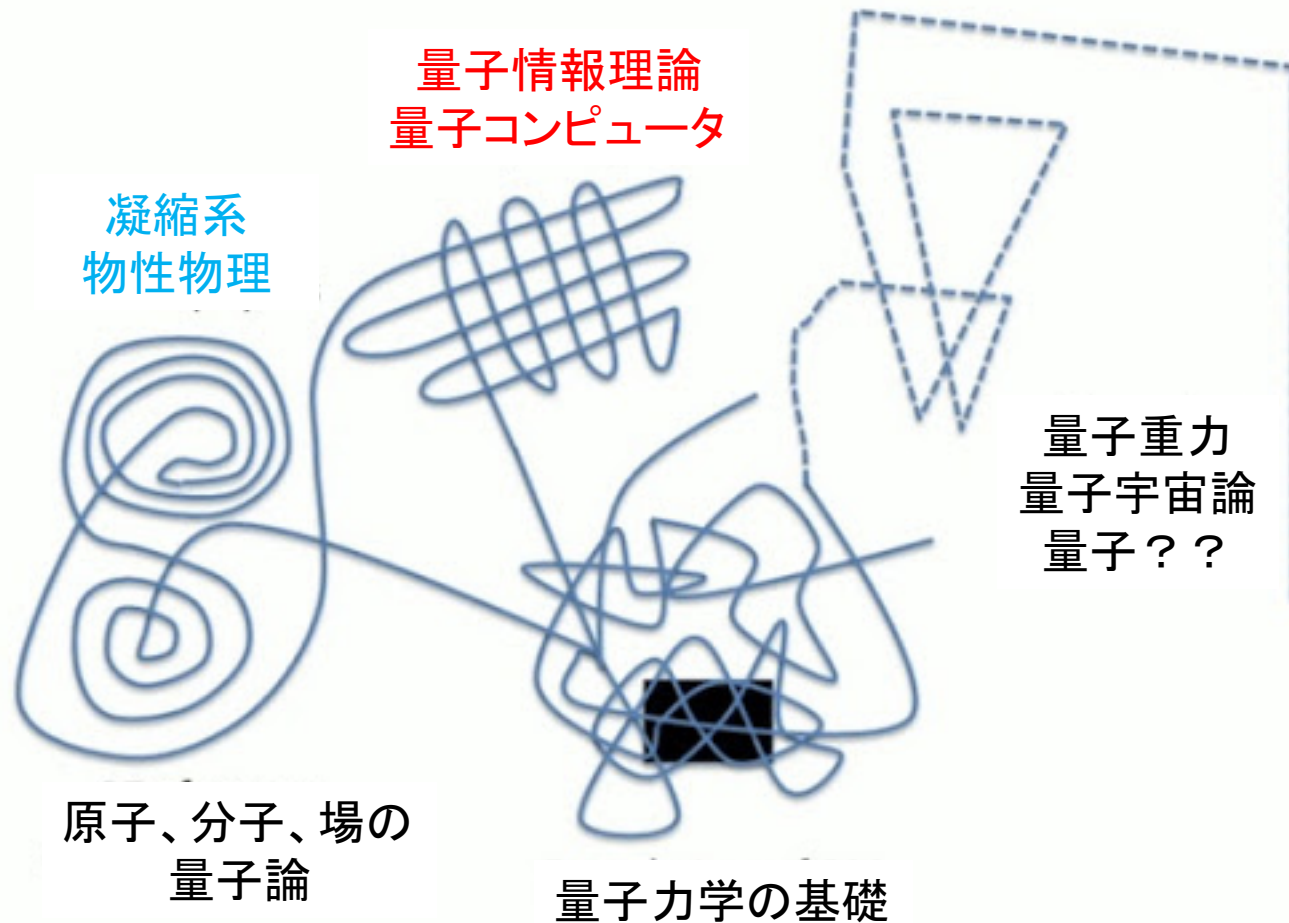
ここでは、現代の物理学者が、現代物理学の課題を、どのように捉えているのかを見ておこう。

現代物理学の俯瞰図 by Susskind



Susskind "Entanglement and Complexity: Gravity and Quantum Mechanics"
<https://goo.gl/J0wSkf>

エンタングルメントの歴史の絵解き



"Our Quantum Society: Living with Entanglement" <https://goo.gl/aWtAzi>
Cathryn Carson <https://goo.gl/j7bE57>

量子多体系の物理

(場の理論、物性理論、統計力学、可解系)

AdS/CFT
(ホログラフィー)



重力理論
(超弦理論)

量子エンタングルメント
テンソルネットワーク

量子情報理論
(情報幾何)



BH情報問題

高柳匡 「重力理論と量子エンタングルメント」

<https://goo.gl/kRPcNI>

なぜ、今、「情報理論」なのか？

It-From-Qubit Complexity Workshop

Stanford University

SUNetID Login

Search this site...



Stanford | Institute for Theoretical Physics

Topics ▾ People ▾ News Events ▾ Videos About ▾

Events

All Upcoming Events

Events by Series

SITP Mini-Courses

SITP Colloquia

Physics Colloquia

Seminar Schedule

Past Events

It-From-Qubit Complexity Workshop

MARCH 20, 2017 (ALL DAY) TO MARCH 22, 2017 (ALL DAY)

SITP OFFICES

REGISTRATION REQUIRED.

Computational Complexity meets Quantum Gravity.

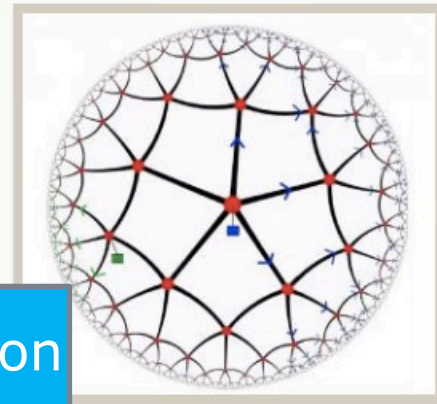
More Information:

It-From-Qubit Complexity Workshop

QUANTUM GRAVITY AND BLACK HOLES

QUANTUM INFORMATION

Quantum Information



2017/03/20 - 22

<https://sitp.stanford.edu/events/it-qubit-complexity-workshop>

科学によって切り開かれる時代

- 科学は、現実的で実践的な技術の基礎づけを目指しているわけではない。むしろ、科学による新しい自然認識の獲得こそが、新しい技術にとって不可欠の土台を提供している。
 - 17世紀の、ニュートン力学の成立を頂点とする、いわゆる「科学革命」は、中世の自然観・社会観と決別し、18世紀の産業革命を準備した。
 - 19世紀のマックスウェルによる電磁気学の成立は、電気の時代の扉を開いた。また、20世紀の相対論と量子論の成立、分子生物学の登場は、20世紀の科学と技術に、深い影響を与えた。
-

なぜ、今、「情報理論」なのか？

- 現代科学の世界で、「情報＝エントロピー」をキー・コンセプトとして、相対論と量子論の統一をはじめとして、科学の基礎の再構成が、急速に進行中だからである。20世紀の相対論・量子論の登場による自然観の変化を上回るインパクトを持つ、新しい自然観が生まれつつある。
 - これらの変化が、物理の基礎理論の問題にとどまらず、技術とつながりが深い、「物性論」の領域を舞台として展開されていることも、重要である。科学の世界の変化は、以前より、短いリードタイムで、新しい技術を準備していこう。
 - 具体的には、様々な量子デバイスと超電導マテリアルが、21世紀の情報とエネルギーの世界を大きく変えるだろう。
-

Appendix

- 複雑さについて
- ディープ・ラーニングとエントロピー



複雑さについて

コルモゴロフの複雑性

シャノンの情報量のもどかしさ

- シャノンの情報量は、ある確率分布に対して与えられる。(他のエントロピー概念でもそうだ。)
 - それは、特定のメッセージ、例えば、一つのメッセージの持つ情報については、何も語ってくれない。それは、言葉の「意味」を直接与えるものではない。
-

Compositionality

- compositionality というのは、複雑なものが単純なものから構成されるという性質のこと。
 - 巨大なコンピュータのプログラムも、それを構成しているのは、単純なstatementの集まりなので、こうした性質を持つ。また、僕が今書いている文章も、分解すれば単純な語の集まりなので、こうした性質を持つ。
 - プログラムでも自然言語でも、単純なものから複雑なものを構成するルールは、文法=syntaxとして与えられるのだが、我々が複雑なものを「把握」するスタイルは、必ずしもsyntaxの総和に還元されない。
 - もっとも、こう言い切るのが妥当かは微妙なところで、これらの例では、syntaxなしでは複雑なものは構成できないのである。
-

「意味」による複雑なものの理解

- 結論を先取りすれば、我々は、複雑なものを、その「意味 = semantics」で把握しようとする。
- 自然言語の場合、それは明らかだと思う。母語の場合、我々は、文法を意識することはない。それは、無意識のうちに自動的に処理され、我々は、その意味を、直接に(と言っても、それは、文法的なcompositionの結果なので、その意味では間接的に)、理解する。
- プログラムでも、我々にとって重要なのは、それを構成する個々のコードの断片ではなく、その全体としての「働き」である。プログラムでも「意味論」は重要になる。いくつかの試みがある。
- compositionality(構成性)は、reducibility(還元可能性)に、関連しているのだが、単純に、ベクトルの向きを変えれば二つの概念は等しくなるというわけでもなさそうだ。

compositionalityとエントロピー

- 無数の分子からなる(と言っても有限個なのだが)ガスの振る舞いを考えよう。そこでは、この系の持つ「情報」としては、個々の分子についてのミクロな情報ではなく、「体積・温度・圧力・エントロピー」といったマクロな情報が有用である。
 - こうした統計力学的アプローチは、compositionalityを満たす系に一つのsemanticsを与えようとするものと考えることができる。もちろん、量子力学的系についても、そうしたアプローチは可能になる。
 - 僕は、意味の形式的理論は、LawvereのFunctor Semanticsで十分だと考えていたキライがあるのだが、どうも、もっと色々考えるべきことはあるみたいだ。
-

コルモゴロフの複雑性

- 計算機科学の分野では、「複雑さ」の尺度に、「コルモゴロフの複雑性」という概念がよく用いられる。それは、ある対象の「複雑さ」を、その対象を出力するプログラムの「大きさ」に置き換えて説明しようというアイデアに基づいている。
- 簡単なものを出力するプログラムは短く、複雑なものを出力するプログラムは長い。これは、プログラマーなら経験的・直感的に理解できることだと思う。
- 同じ結果を返すいくらでも冗長なプログラムはかけるのだから、出力の「複雑さ」を、プログラムの「大きさ」に対応づけようとするのなら、その結果を返す「最も短いプログラム」の大きさを考えるのがいいことになる。
- ある対象の「コルモゴロフの複雑性」と言うのは、その対象を出力する「最も短いプログラム」の「大きさ」のことである。

コルモゴロフの複雑性は存在する

- 繰り返しになるが、ある対象の「コルモゴロフの複雑性」と言うのは、その対象を出力するプログラムの大きさが最小の自然数である。ある対象を出力する、同じ機能を持つ多数のプログラムがあったとしても、そのプログラム・サイズからなる自然数の全体の集まりを考えれば、その中に、最小の自然数が、一つは存在するはずである。
- 実行環境を固定しても、実際には、ある対象の「コルモゴロフの複雑性」を具体的に「計算」することは難しいかもしれない。ただ、ここは「コルモゴロフの複雑性」理解の上で、飛躍が必要なポイントでもある。
- 重要なことは、具体的にその数字を示すことができなくとも、また具体的にそのプログラムを示すことができなくとも、最小のプログラム・サイズが存在するのは確かであるということである。

チャイティンの不完全性定理

- コルモゴロフは、旧ソ連の数学者で、「コルモゴロフの複雑性」が定式化されたのは、1960年代のこと。今から50年近く前の話だ。「コルモゴロフの複雑性」が奇妙な性質を持つことを、驚くような形でまとめたのは、チャイティン(G. J. Chaitin)である。
 - 「どのような特定のビット列をとっても、そのコルモゴロフの複雑性が L 以上であることを証明できないような数 L が存在する。」これを「チャイティンの不完全性定理」という。
 - いくらでも複雑なものは存在する。どんな数に対しても、その数以上のコルモゴロフ複雑性を持つビット列は、無限に存在することを、我々は証明できる。ただ、特定のビット列が与えられた時、そのコルモゴロフ複雑性が L 以上であることを証明できない L が存在すると言うのだ。
-

「複雑さ」には、限界がある

- どういうことか？
 - いくらでも複雑なものを、我々は、無限に考えることができる。複雑なビット列を考えて、それとは違うもっと複雑なビット列を考えて、もっと考えて、..... これを無限に繰り返しても、具体的なビット列を考える限り、その複雑さは L を超えないと言うのだ。
 - まるで、孫悟空がいくら飛び回っても、お釈迦様の手のひらから飛び出すことができないように。複雑さを求めて、宇宙の果てまで来たつもりだったのに、そこには超えられない壁 L があることを見つけたと言うことだ。
 - 「複雑さ」(我々が具体的に考える)には、乗り越えられない壁 L があるのだ。
-

「複雑さ」には、限界がある

- いくらでも複雑なものを、我々は、無限に考えることができる。複雑なビット列を考えて、それとは違うもっと複雑なビット列を考えて、もっと考えて、..... これを無限に繰り返しても、具体的なビット列を考える限り、その複雑さは L を超えないと言うのだ。
 - まるで、孫悟空がいくら飛び回っても、お釈迦様の手のひらから飛び出すことができないように。複雑さを求めて、宇宙の果てまで来たつもりだったのに、そこには超えられない壁 L があることを見つけたと言うことだ。
 - 「複雑さ」(我々が具体的に考えうる)には、乗り越えられない壁 L があるのだ。
-

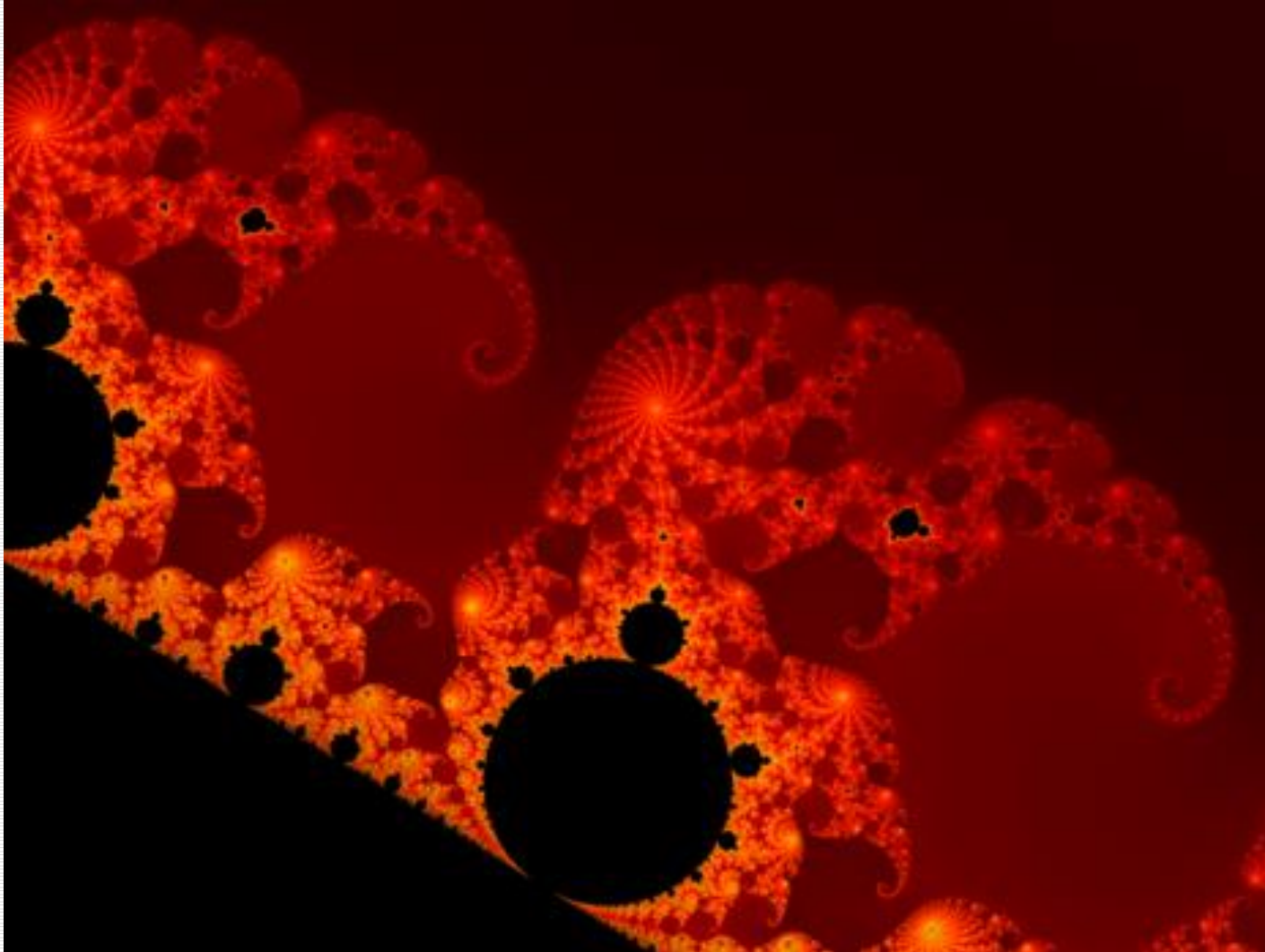
複雑さの上限Lは、驚くほど小さい！

- そういうと、Lは、とんでもなく巨大な数だろうと考えたくなるのだが、Lは、驚くほど小さいらしい。
- compositionality というのは、複雑なものが単純なものから構成されるという性質のことだと言ったが、単純なものから複雑なものを構成するルールをが、Nビットで記述できるとすると、複雑さの壁 L は、
$$L \leq N + 2359$$
程度だという。正確に、Lの値を求めることはできないが、あるコンピューター科学者は、数キロバイトだと推測している。
- 小さい！ 特定のビット列を出力するプログラムは、そのビット列がどんなに複雑でも、数キロバイトに収まるということで、皆さんが普段作っているプログラムが、数キロバイトになるはずだということではないので、ご安心を。

チャイテンの不完全性定理の言い換え

- チャイティンは、こういう言い方もしている。
 - 「あるプログラムを、それより小さいサイズのプログラムが同じ出力を行うことがない時、"エレガント"と呼ぶことにしよう。あるプログラムが "エレガント" であることを、我々は証明できない。」
-

小さなプログラムでも、 複雑なビット列を生み出すことができることの例



この画像は、1ピクセルあたり24bitカラーを用いていて、全体で、162万bitあるのだが、これを生成するプログラムは、本当に小さい。

DemoSceneのコンテストで優勝した動画。見知らぬ惑星の描写だが、音楽含めて、プログラム・サイズは、たったの4Kだという。信じられない！

<https://www.youtube.com/watch?v=I5CTFMuFvb0>

$P \neq NP$ 問題

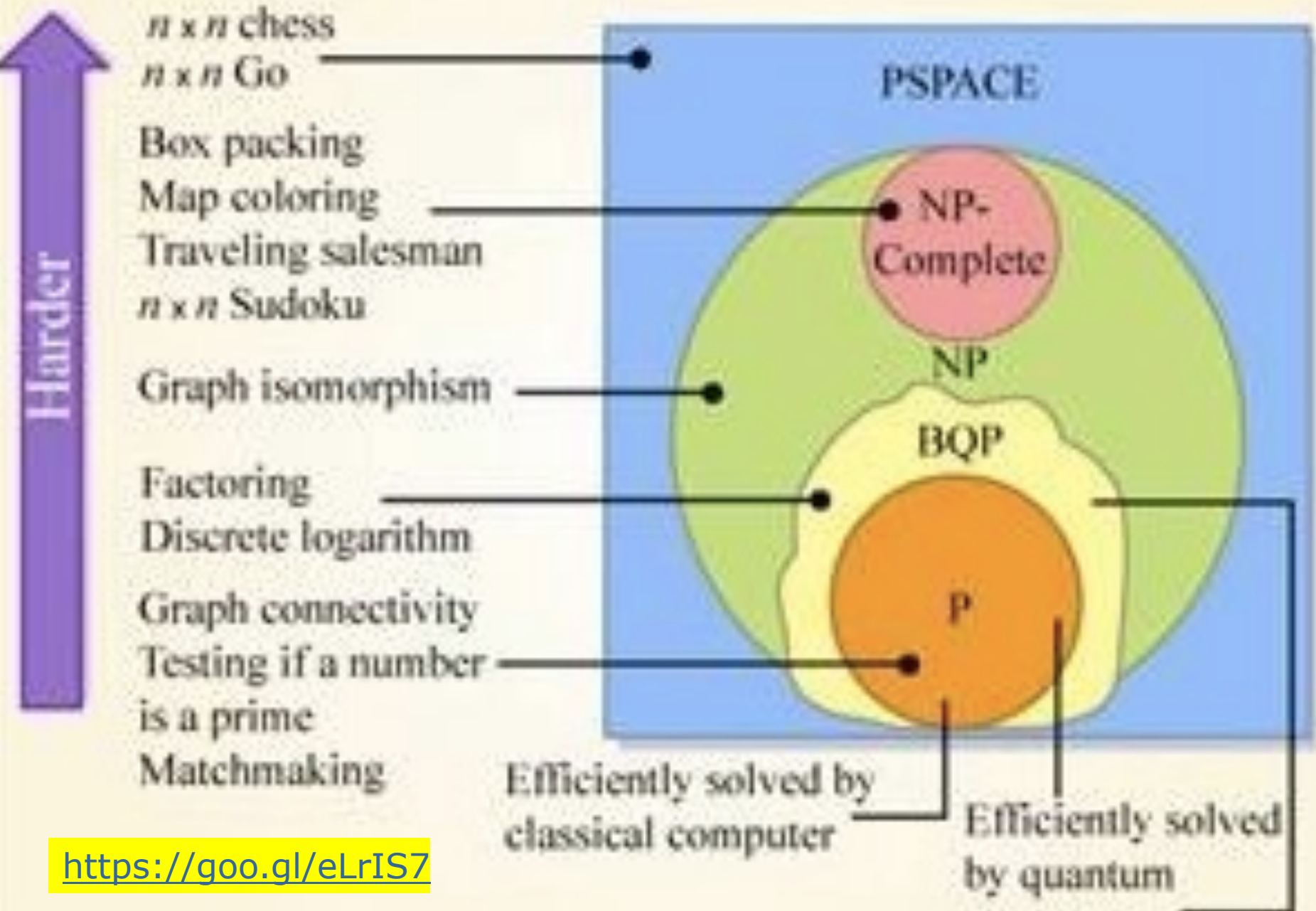
P≠NP問題

- P≠NP問題は、Nashが提起した計算の複雑性の階層についての代表的な問題の一つである。計算の複雑性の理論は、古典的には、1930年代のChurch-Turingの「計算可能性」の議論に起源を持つ。
- 1985年にDeutschは、その定式化を、量子コンピュータを含む全ての物理的な機械の計算可能性に拡大した。Church-Turing-Deutsch principle。ここでの議論は、現代の量子コンピュータの出発点になった。（アイデアは、Feynman が提供した。）

PとNP

- ❑ 多項式時間で計算できる複雑さを、Polynomial-Timeの頭文字をとってPで表すのだが、NPというのは、Non-Polynomial-Time の略ではない。Nondeterministic Polynomial-Timeの略だ。
- ❑ NPを多項式時間では計算できない複雑さだと、誤解している人も多いのだが、そもそも、このNPのネーミングが、よくないのだと思う。
- ❑ 次の図を見れば、Pの外側全体(多項式時間で計算できないもの)が、NPにはなっていないことがわかるだろう。ある問題が、正しいとわかっている時に、その検証が多項式時間でできるものをNPと呼ぶ。
- ❑ ある数が、二つの素数の積に分解できることを見つけるのは難しいが、二つの素数の積がある数になることは、簡単に(多項式時間で)計算できる。だから、素因数分解の問題は、Pかどうかはわからないが、NP問題である。(実は、素因数分解が、多項式時間で計算可能なことは、最近になってわかった。)

Example Problems



-
- 多項式時間で計算できない数には、とんでもなく複雑な数がある。今年の4月に、Adam Yedidia と Scott Aaronsonは、BB(7910)が、普通の集合論の枠組みの中では決定できないことを証明した。BBは、Busy Beaver数という、複雑だが、きちんと定義された、しかも有限の数だ。
 - 要するに、「有限」の値を持つのだが、どういう数学的な手段を持ちいても、我々が知り得ない数字があるということ。さらに、その限界が、7910とか1919といった、比較的小さな数字で特徴付けられるというのだ。
 - チェイティンの不完全性定理の示す、我々人間が(機械でも同じことだ)複雑さにチャレンジする時の、「複雑さの壁」(しかも、小さな数字で特徴付けられる)と同じような問題が、計算量の理論の中でも見つかったということだと思う。

-
- ビーバー数についてのトピックは、以前の僕の次の二つの投稿を参考にしてほしい。
 - 「Busy Beaver Game問題について」<https://goo.gl/3QLsMc>
 - 「有限だが計算できない数の発見について」
<https://goo.gl/oV2b1o>
 - どうやっても計算できない数は、PであれNPであれ、計算可能な数より、複雑である。
-

量子論的計算複雑性理論

- 量子論的複雑性理論(Quantum complexity theory)は、古典的な計算複雑性の理論を、量子コンピュータと量子情報理論に拡大したもの。ここでは、古典的なP, NPに対応する量子計算可能性のクラス BQP, QMA が存在する。P≠NP問題の解決が自動的にBQP≠ QMA問題の解決を与えるものではない。(と思う)
 - 計算の複雑さの議論は、もっと先に進んでいる。それは、現代の物理学の深い部分と繋がっている。(P≠NPさえ、証明できていないのになと思う人はいるのかも。)たとえば、これ。
"Recent papers by Susskind and Tao illustrate the long reach of computation"
<http://www.scottaaronson.com/blog/?p=1697>
-

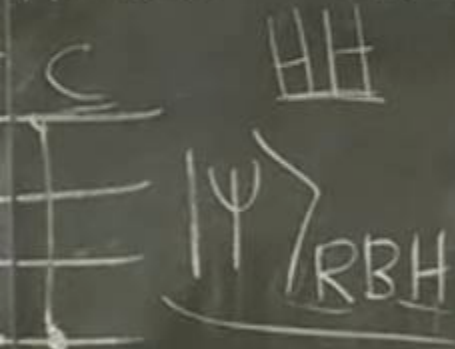
-
- 量子論的複雑性理論の中心人物が、Scott Aaronsonである。
(今日も、彼のblogは更新されていない。自分が追いかけてきた問題が、他の人に解決されるのを見るのは、悔しいのかもしれない。でも、コメントが欲しい。)
 - ここ数年、量子Entanglementのエントロピーの研究が急速に進んで、物理学の中で複雑性理論への関心が高まっている。関心の中心になったのは、ブラックホールでの情報の在り方である。
-

-
- 写真は、Raamsdonkの講演で、自分のシャツをアピールする物理学者 Susskind。やんちゃである。胸には、"I love Complexity" と書かれている。Mark van Raamsdonk "Gravity and Entanglement" <https://goo.gl/jd9dc4>
 - もう一つの写真は、ブラックホール内部での「情報」についての Harlow-Haydenの驚くべき議論を紹介するScott Aaronson。"Computational complexity underpinnings of the Harlow-Hayden argument" <https://goo.gl/EEr4JU>
-





HAYDEN ARGUMENT



Size (#gates) $\leq n^{O(1)}$

$$\rho_{BH} = I_{BH}$$

$\exists$ 1 qubit of distillable entanglement between R & B

Alice's task

Apply U_R to R so that



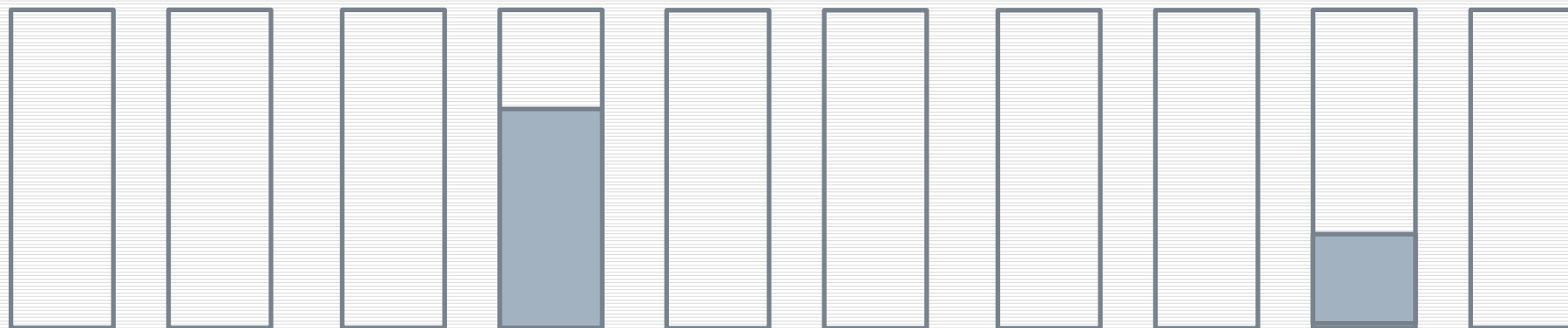
ディープ・ラーニングとエントロピー

SoftMax関数

softmax 関数

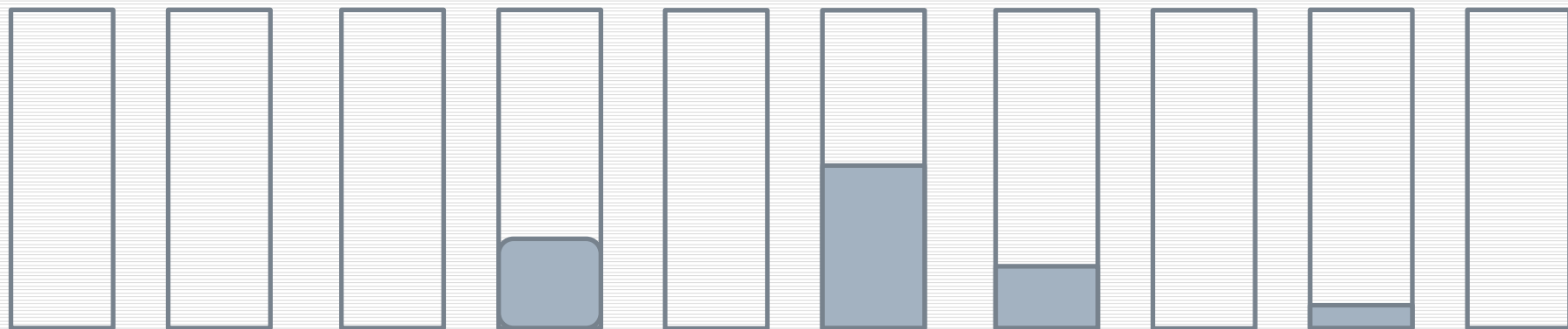
- クラス分けの場合の出力に利用されるactivator。
手書き数字画像の認識だと、提示された画像が0から9までの10個の数字のどれかだと判断しなければならない。
 - ニューラル・ネットワークでは、こうした場合、出力層に、0から9までの10個の数字に対応した10個のノードが並び、それぞれのノードの値が、「提示された画像がその数字である確率」であるようにする。
 - ノードごとに、確率を計算するのが softmax関数である。最終的には、10個のノードの中で、その確率が最大であるものが選ばれる。
-

softmax 関数の出力例



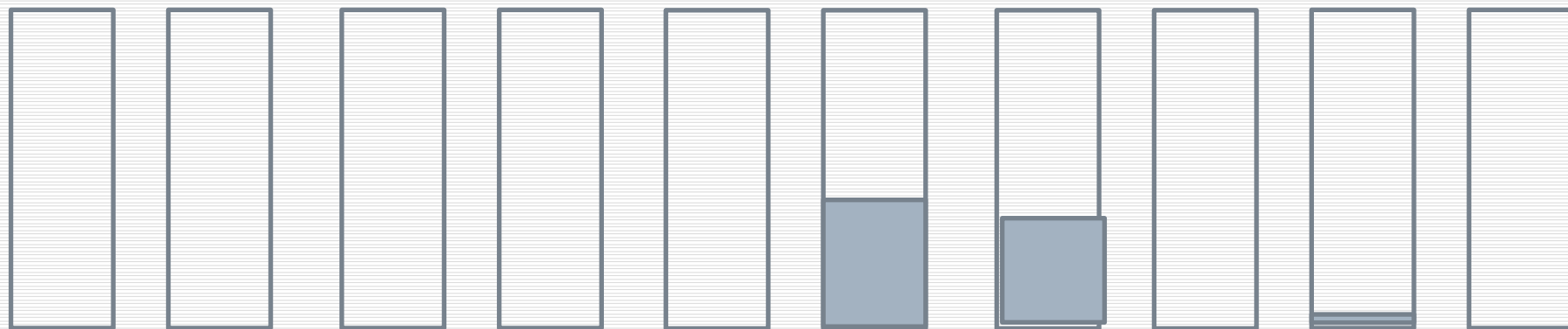
ノード0 ノード1 ノード2 ノード3 ノード4 ノード5 ノード6 ノード7 ノード8 ノード9

softmax 関数の出力例



ノード0 ノード1 ノード2 ノード3 ノード4 ノード5 ノード6 ノード7 ノード8 ノード9

softmax 関数の出力例



ノード0 ノード1 ノード2 ノード3 ノード4 ノード5 ノード6 ノード7 ノード8 ノード9

softmax 関数の性質

- softmax関数は、それが適用される層のノードの softmax関数の値をすべて足し合わせると1になるという性質を持つ。それは、softmaxの出力が「確率」であることを考えれば、すぐわかる。
- 例えば、先の例で、ある画像が3である確率が 0.75であったとしよう。そのことは、その画像が3以外の数字である確率は、 $1 - 0.75 = 0.25$ だということである。3以外のノードの確率(softmaxの出力の値)を全部足しても、0.25にしかないということである。
- 次のこともわかる。あるノードのsoftmaxの値が上がれば、他のノードの値は下がる。あるノードの値を下げれば、他のノードの値は上がる。全部足して1になるように、値は変化する。

One-Hot-Vector

- 先のsoftmax関数の出力例の、一番上のノードは、「この画像は数字nである」という、ニューラル・ネットワークの判断を表しているの。もちろん、softmaxの値が一番大きいところ選ばれて、そこに一つだけに1が入っていて、残りはすべて0になっている。
- こうした一つだけに1が、残りすべてが0のベクトルを、One-Hot-Vector と呼ぶ。
- One-Hot-Vectorは、クtas分けの出力に現れるだけではない。ニューラル・ネットワークを「学習」させるための訓練用データは、画像と一緒に「この画像は数字nである」という「正解」の情報を持っていなければならない。訓練用データは、画像と正解を示すOne-Hot-Vectorの二種類のデータから構成されている。

tensorflow/tensorflow/python/ops/nn_ops.py

https://github.com/tensorflow/tensorflow/blob/r1.3/tensorflow/python/ops/nn_ops.py

softmax (logits, dim=-1, name=None)

$$p_i = \frac{e^{\beta E_i}}{\sum_j e^{\beta E_j}}$$

For each batch `i` and class `j` we have

```
softmax = exp(logits) / reduce_sum(exp(logits), dim)
```



Partition Function

Args:

logits: A non-empty `Tensor`. Must be one of the following types: `half`, `float32`, `float64`.

dim: The dimension softmax would be performed on. The default is -1 which indicates the last dimension.

name: A name for the operation (optional).

Returns: A `Tensor`. Has the same type as `logits`. Same shape as `logits`.

return _softmax(logits, gen_nn_ops._softmax, dim, name)

softmaxは、_softmaxを呼んでいる。

_softmax (logits, compute_op, dim=-1, name=None)

"""Helper function for softmax and log_softmax.

It reshapes and transposes the input logits into a 2-D Tensor and then invokes the `tf.nn._softmax` or `tf.nn._log_softmax` function. The output would be transposed and reshaped back.

Args:

logits: A non-empty ``Tensor``. Must be one of the following types: ``half``, ``float32``, ``float64``.

compute_op: Either `gen_nn_ops._softmax` or `gen_nn_ops._log_softmax`

dim: The dimension softmax would be performed on. The default is -1 which indicates the last dimension.

name: A name for the operation (optional).

Returns: A ``Tensor``. Has the same type as ``logits``. Same shape as ``logits``.

softmax_cross_entropy_with_logits

""Computes softmax cross entropy between `logits` and `labels`. Measures the probability error in discrete classification tasks in which the classes are mutually exclusive (each entry is in exactly one class). For example, each CIFAR-10 image is labeled with one and only one label: an image can be a dog or a truck, but not both.

****NOTE:**** While the classes are mutually exclusive, their probabilities need not be. All that is required is that each row of `labels` is a valid probability distribution. If they are not, the computation of the gradient will be incorrect. If using exclusive `labels` (wherein one and only one class is true at a time), see `sparse\_softmax\_cross\_entropy\_with\_logits`.

****WARNING:**** This op expects unscaled logits, since it performs a `softmax` on `logits` internally for efficiency. Do not call this op with the output of `softmax`, as it will produce incorrect results.

softmax_cross_entropy_with_logits

```
cost, unused_backprop =  
gen_nn_ops._softmax_cross_entropy_with_logits(  
    precise_logits, labels, name=name)
```

sparse_softmax_cross_entropy_with_logits

""Computes sparse softmax cross entropy between `logits` and `labels`.

Measures the probability error in discrete classification tasks in which the classes are mutually exclusive (each entry is in exactly one class). For example, each CIFAR-10 image is labeled with one and only one label: an image can be a dog or a truck, but not both.

****NOTE:**** For this operation, the probability of a given label is considered exclusive. That is, soft classes are not allowed, and the `labels` vector must provide a single specific index for the true class for each row of `logits` (each minibatch entry). For soft softmax classification with a probability distribution for each entry, see `softmax\_cross\_entropy\_with\_logits`.

Cross Entropy 関数

クロス・エントロピー $H(p,q)$ の定義

The diagram illustrates the definition of cross-entropy $H(p,q)$. The formula is presented within a yellow rounded rectangle:
$$-\sum_N \sum_x p(x) \log(q(x))$$
 Annotations with arrows point to specific parts of the formula:

- マイナス** (Minus) points to the negative sign $-$.
- 目標の確率** (Target probability) points to $p(x)$.
- データの確率** (Data probability) points to $q(x)$.
- サンプルの数** (Number of samples) points to the summation index N .
- 分類の数** (Number of classes) points to the summation index x .

クロス・エントロピー $H(p,q)$ の定義

マイナス

目標の確率

データの確率

$$-\sum_N \sum_x p(x) \log(q(x))$$

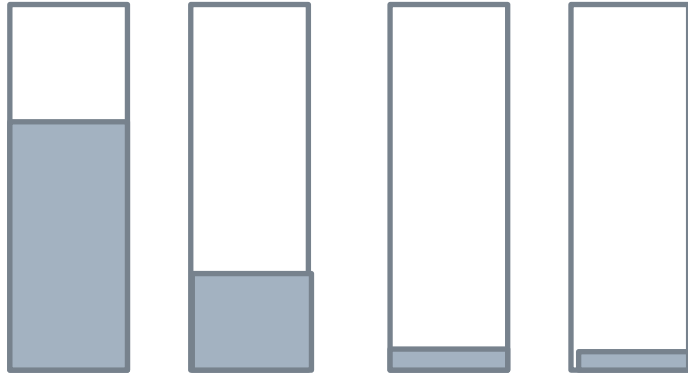
サンプルの数

分類の数

まず、この部分の計算の
サンプルを見ておこう。

$p(x)$

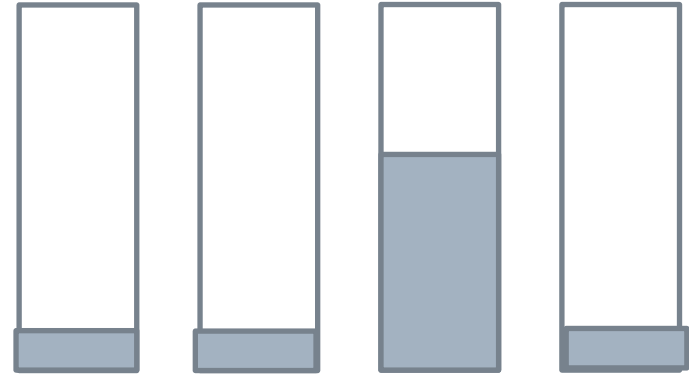
期待される
目標の分布



$$P(1)=2/3, p(2)=1/4, p(3)=5/24, p(4)=5/24$$

$q(x)$

実際の
データの分布



$$q(1)=1/8, q(2)=1/8, q(3)=5/8, q(4)=1/8$$

$$x=1 \text{ の時 } 2/3 \log(1/8) = -2/3 \log 8 = -2 \log 2$$

$$x=2 \text{ の時 } 1/4 \log(1/8) = -1/4 \log(8) = -3/4 \log 2$$

$$x=3 \text{ の時 } 5/24 \log(5/8) = 5/24 \log(5) - 15/24 \log 2$$

$$x=4 \text{ の時 } 5/24 \log(1/8) = -5/24 \log(8) = -15/24 \log 2$$

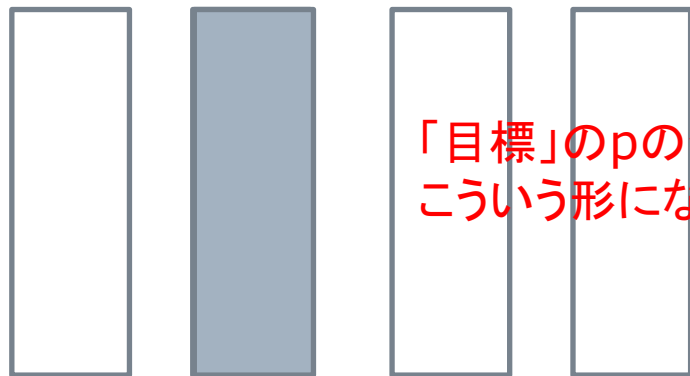
全部足し合わせると

$$-(48 \log 2 + 18 \log 2 - 5 \log 5 + 15 \log 2 + 15 \log 2) / 24 = -(96 \log 2 - 5 \log 5) / 24$$

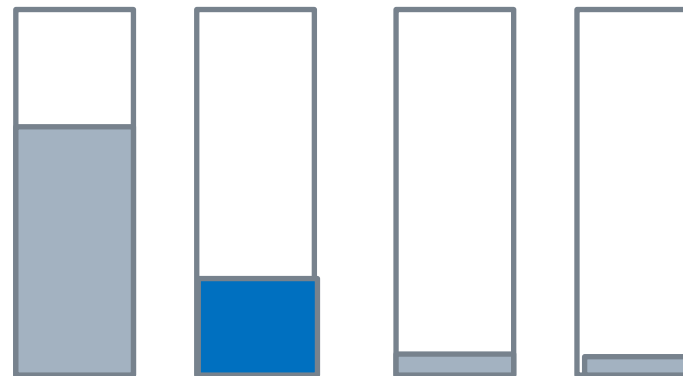
$$H(p, q) = (96 \log 2 - 5 \log 5) / 24$$

$$\sum_x p(x) \log(q(x))$$

の計算

H_1 $p(x)$ 期待される
目標の分布「目標」のpの分布は、
こういう形になる。

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

 $q(x)$ 実際の
データの分布

$$q(1)=2/3, q(2)=1/4, q(3)=5/24, q(4)=5/24$$

$$x=1 \text{ の時 } 0 \log 2/3$$

$$x=2 \text{ の時 } 1 \log 1/4$$

$$x=3 \text{ の時 } 0 \log 5/24$$

$$x=4 \text{ の時 } 0 \log 5/24$$

全部足して

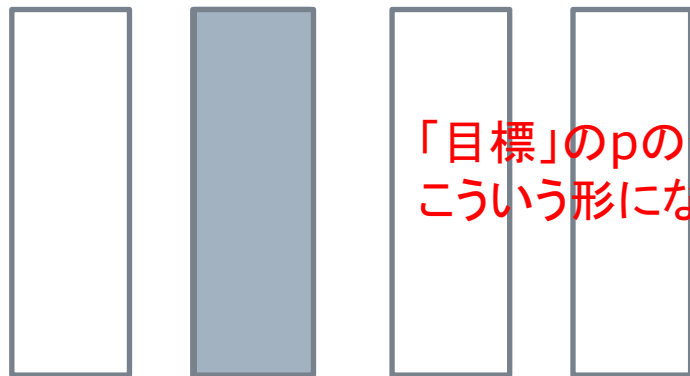
$$\log 1/4 = -\log 4 = -2 \log 2$$

$$H_1(p, q) = 2 \log 2$$

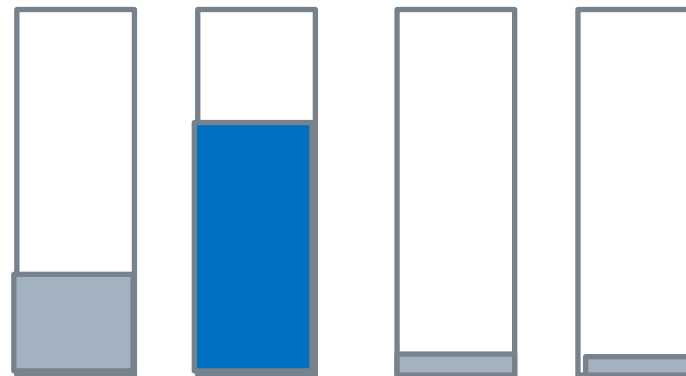
 $p(2) = 1$ の場合の

$$\sum_x p(x) \log(q(x))$$

の計算

H_2 $p(x)$ 期待される
目標の分布「目標」のpの分布は、
こういう形になる。

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

 $q(x)$ 実際の
データの分布

$$q(1)=1/4, q(2)=2/3, q(3)=5/24, q(4)=5/24$$

$$x=1 \text{ の時 } 0 \log 1/4$$

$$x=2 \text{ の時 } 1 \log 2/3$$

$$x=3 \text{ の時 } 0 \log 5/24$$

$$x=4 \text{ の時 } 0 \log 5/24$$

全部で

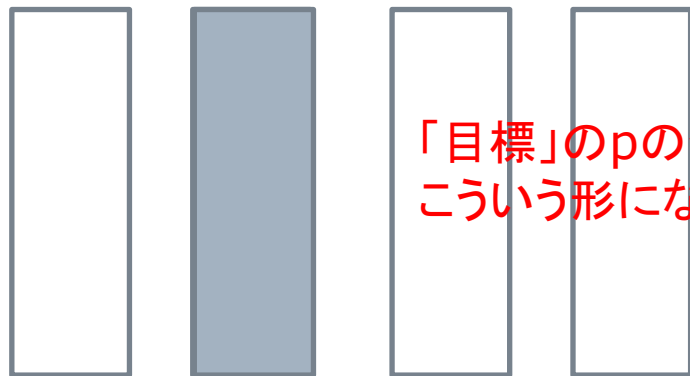
$$\log 2/3 = \log 2 - \log 3$$

$$H_2(p, q) = \log 3 - \log 2$$

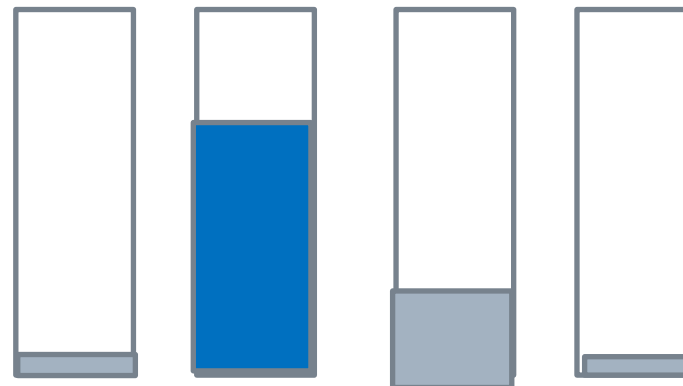
 $p(2) = 1$ の場合の

$$\sum_x p(x) \log(q(x))$$

の計算

H_3 $p(x)$ 期待される
目標の分布「目標」のpの分布は、
こういう形になる。

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

 $q(x)$ 実際の
データの分布

$$q(1)=5/24, q(2)=2/3, q(3)=1/4, q(4)=5/24$$

$$x=1 \text{ の時 } 0 \log 5/24$$

$$x=2 \text{ の時 } 1 \log 2/3$$

$$x=3 \text{ の時 } 0 \log 1/4$$

$$x=4 \text{ の時 } 0 \log 5/24$$

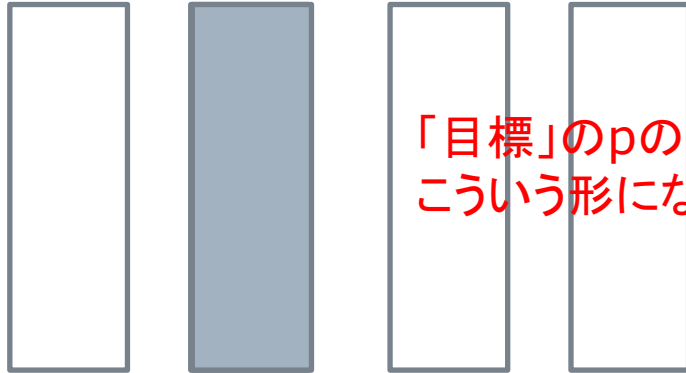
全部足し合わせると
 $\log 2/3$

$$H_3(p, q) = \log 3 - \log 2$$

 $p(2) = 1$ の場合の

$$\sum_x p(x) \log(q(x))$$

の計算

H_4 $p(x)$ 期待される
目標の分布「目標」のpの分布は、
こういう形になる。

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

 $q(x)$ 実際の
データの分布

$$q(1)=0, q(2)=1, q(3)=0, q(4)=0$$

$$x=1 \text{ の時 } 0 \log 0 = 0$$

$$x=2 \text{ の時 } 1 \log 1 = 0$$

$$x=3 \text{ の時 } 0 \log 0 = 0$$

$$x=4 \text{ の時 } 0 \log 0 = 0$$

$$H_5(p, q) = 0$$

 $p(2) = 1$ の場合の

$$\sum_x p(x) \log(q(x))$$

の計算

$p(x)$ が one-hot-value の時

$$H(p, q) = - \sum_N \sum_x p(x) \log(q(x))$$

$$= - \sum_N \log(q(i))$$

= - { サンプル中の正解値をとった i についての $\log(q(i))$ の和 }

$0 \leq q(i) \leq 1$ で、 $\log(q(i)) \leq 0$ だが、

$q(i) \rightarrow 0$ の時 $\log(q(i)) \rightarrow -\infty$

$q(i) \rightarrow 1$ の時 $\log(q(i)) \rightarrow 0$

だから、 $q(i)$ が正解に近いほど、 $H(p, q)$ は小さくなる。

クロス・エントロピー $H(p,q)$ の定義

マイナス 目標の確率 データの確率

↓ ↓ ↓

$$-\sum_N \sum_x p(x) \log(q(x))$$

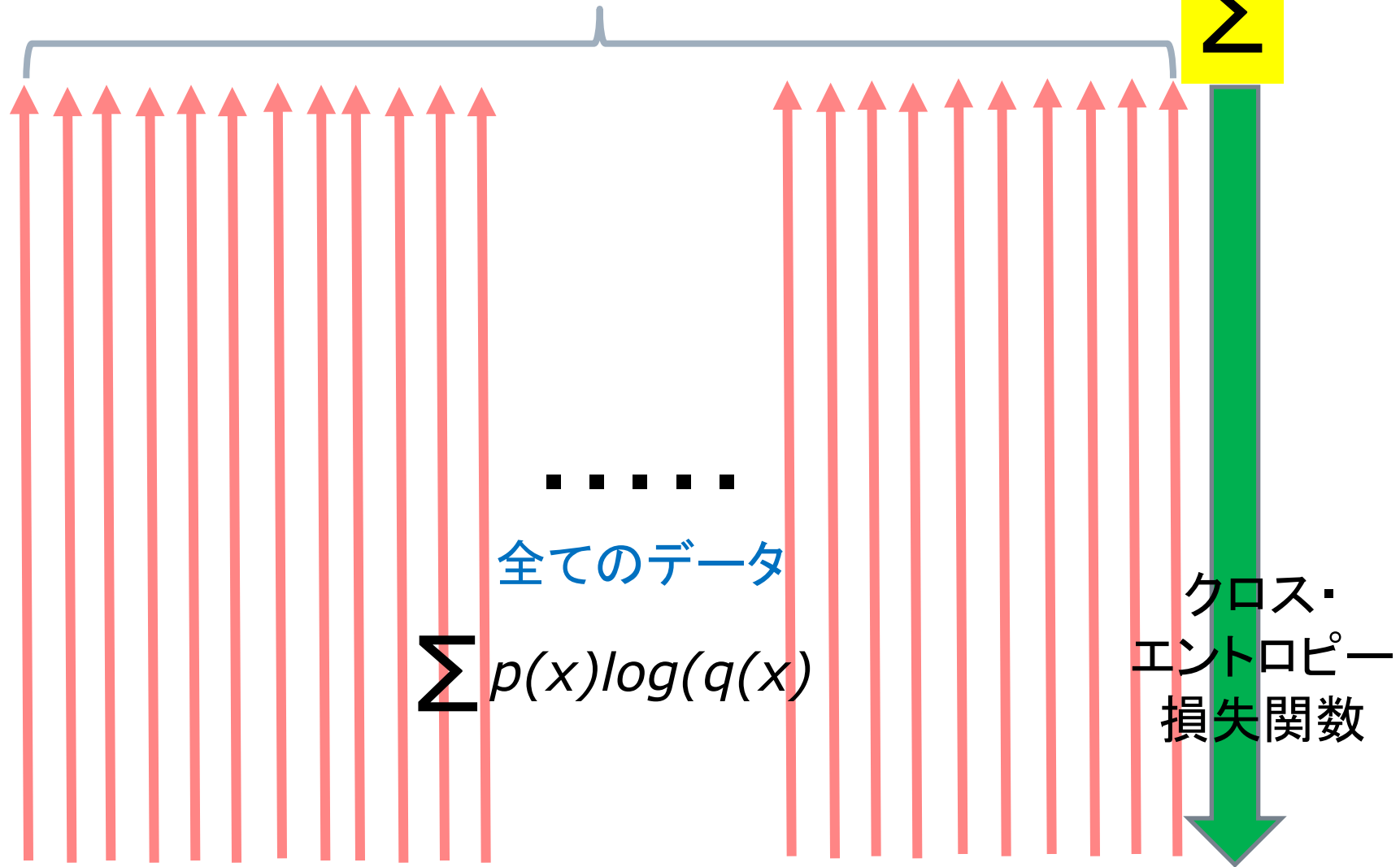
↑ ↑

サンプルの数 分類の数

この部分は、どうなるか？

クロス・エントロピーの値の計算

すべてのサンプル・データ x について、 $\sum p(x)\log(q(x))$ を計算し、
(赤い矢印)、その後、これらをすべて足し合わせる。(緑の矢印)。

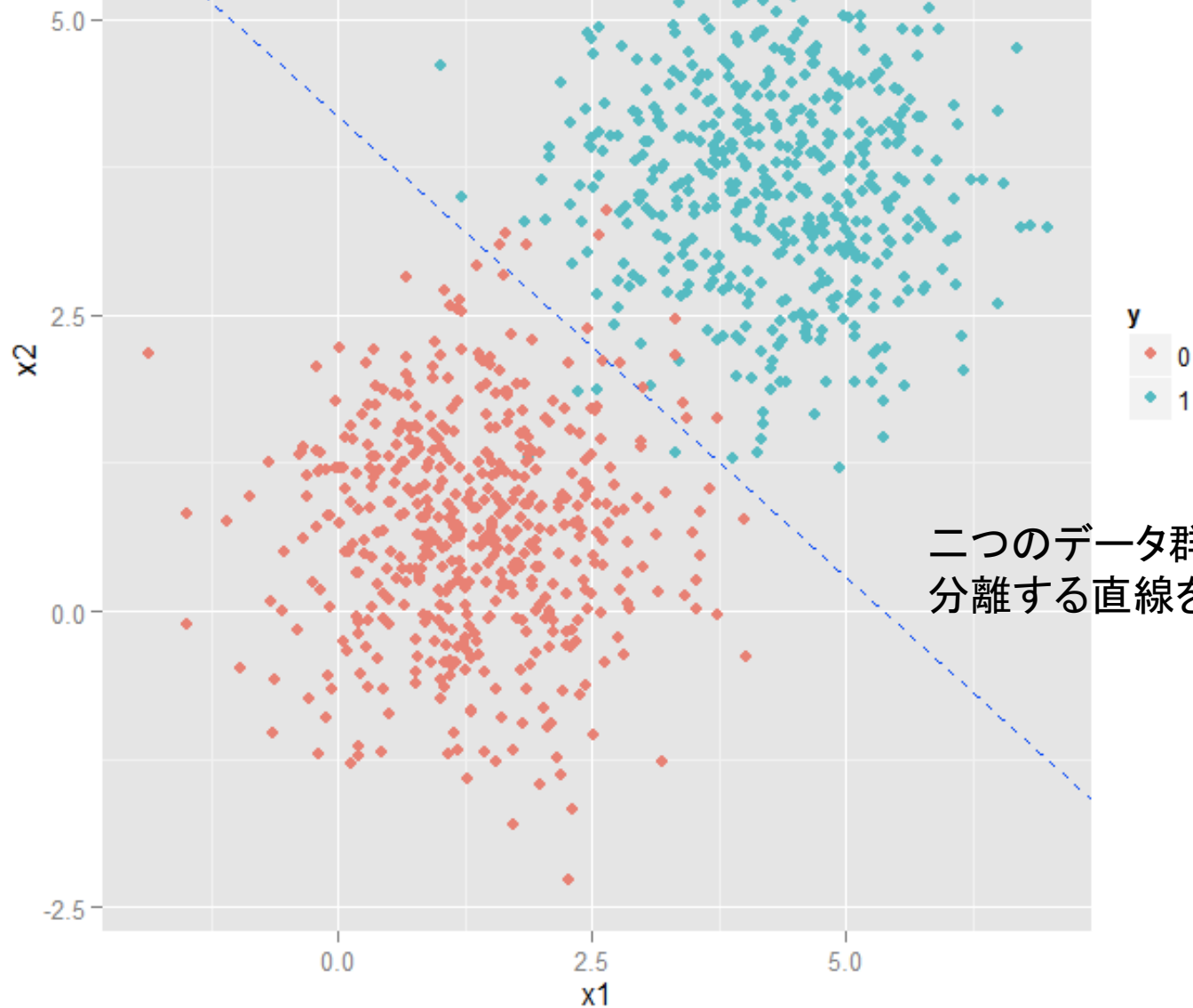


バイナリー・クロス・エントロピー

先に見たクロス・エントロピーは、二つの分布 p と q が「似ているか？」については、あまり有効な情報を与えない。なぜなら、 $p(x)=1$ 以外の x の値での q の値は、すべて無視するので。

ただ、次のように定義を変更すると、二つの分布の「近さ」を数値化できる。(計算が増えるので、必要ないかもしれないのだが)

これは、統計学での
二項分類のLogistic
回帰にあたる。



二つのデータ群を
分離する直線を求める。

$$H(p, q) = - \sum_x \{ p(x) \log(q(x)) + (1-p(x)) \log(1-q(x)) \}$$

$P(x), q(x)$ は、確率を表しているので

$$0 \leq p(x) \leq 1, 0 \leq q(x) \leq 1$$

$$0 \leq 1-p(x) \leq 1, 0 \leq 1-q(x) \leq 1$$

$P(x) \geq 0$ で、 $\log(q(x)) \leq 0$ だから

$$p(x) \log(q(x)) \leq 0$$

$(1-p(x)) \geq 0$ で、 $\log(1-q(x)) \leq 0$ だから

$$(1-p(x)) \log(1-q(x)) \leq 0$$

よって、 Σ の中の

$$p(x) \log(q(x)) + (1-p(x)) \log(1-q(x)) \leq 0 \text{ となる}$$

Σ の前に マイナス符号が付いているので、 **$H(p, q) \geq 0$**

Binary Entoropy

の式の特徴

マイナスが
付いている

p(x)が成り立つ確率

p(x)が成り立たない確率

$$H(p, q) = - \sum_x \{ p(x) \log(q(x)) + (1-p(x)) \log(1-q(x)) \}$$

目標とする
確率

データの
確率

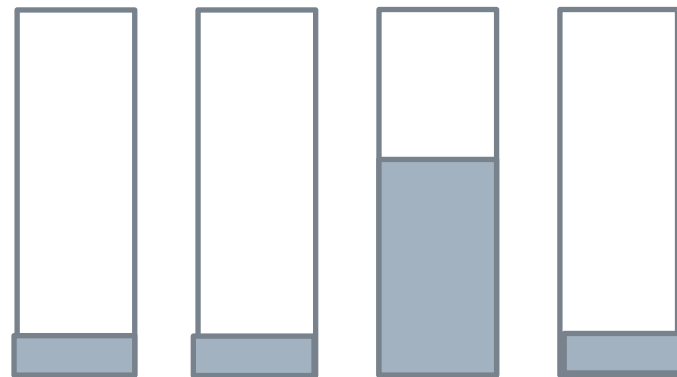
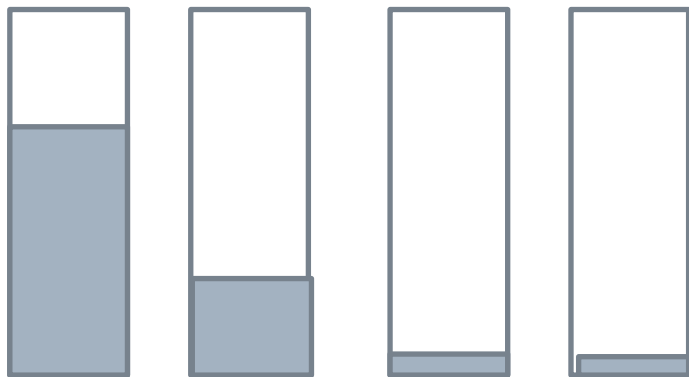
p(x)=0の時は、
この項が消える

p(x)=1の時は、
この項が消える

$$H(p, q) = - \sum_x \{ [p(x) \log(q(x))] + [(1-p(x)) \log(1-q(x))] \}$$

$$H(p) = - \sum_x p(x) \log(p(x))$$

ShannonのEntoropy

$p(x)$ $q(x)$ 

$$P(1)=2/3, p(2)=1/4, p(3)=5/24, p(4)=5/24 \quad q(1)=1/8, q(2)=1/8, q(3)=5/8, q(4)=1/8$$

$$x=1 \text{ の時} \quad 2/3 \log(1/8) + (1-2/3) \log(1-1/8) = 2/3 \log(1/8) + 1/3 \log(7/8)$$

$$x=2 \text{ の時} \quad 1/4 \log(1/8) + (1-1/4) \log(1-1/8) = 1/4 \log(1/8) + 3/4 \log(7/8)$$

$$x=3 \text{ の時} \quad 5/24 \log(5/8) + (1-5/24) \log(1-5/8) = 5/24 \log(5/6) + 19/24 \log(3/8)$$

$$x=4 \text{ の時} \quad 5/24 \log(1/8) + (1-5/24) \log(1-1/8) = 5/24 \log(1/8) + 19/24 \log(7/8)$$

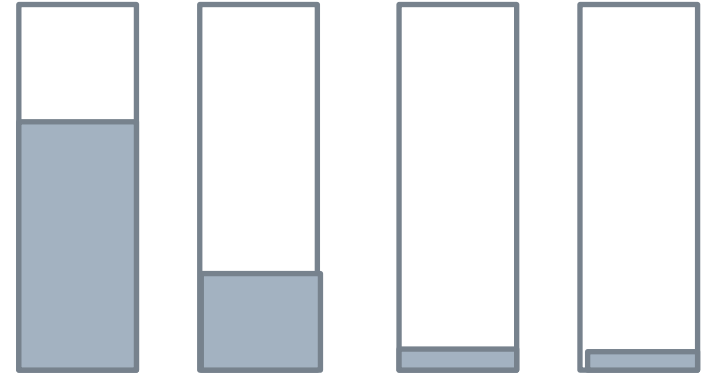
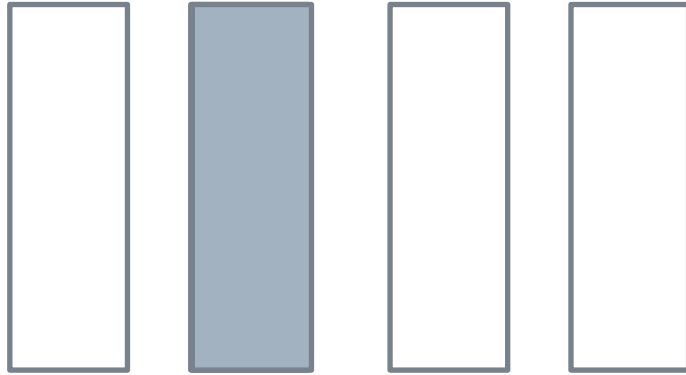
全部足し合わせると

$$\begin{aligned} & (-2/3 \log 8 + 1/3 (\log 7 - \log 8)) + (-1/4 \log 8 + 3/4 (\log 7 - \log 8)) + \\ & (5/24 (\log 5 - \log 6) + 19/24 (\log 3 - \log 8)) + (-5/24 \log 8 + 19/24 (\log 7 - \log 8)) \\ & = (-16 \log 8 + 8 \log 7 - 8 \log 8 - 6 \log 8 + 18 \log 7 - 18 \log 8 + \\ & \quad 5 \log 5 - 5 \log 6 + 19 \log 3 - 19 \log 8 - 5 \log 8 + 19 \log 7 - 19 \log 8) / 24 \\ & = (-(16+8+6+18+19+5+19) \log 8 + (8+18+19) \log 7 - 5 \log 6 + 19 \log 3) / 24 \\ & H(p, q) = - (-91 \log 8 + 45 \log 7 - 5 (\log 2 + \log 3) + 19 \log 3) / 24 \\ & = (278 \log 2 - 45 \log 7 + 14 \log 3) / 24 \end{aligned}$$

H₁

p(x)

q(x)



$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

$$q(1)=2/3, q(2)=1/4, q(3)=5/24, q(4)=5/24$$

$$x=1\text{の時} \quad 0\log 2/3 + 1\log(1-2/3) = \log(1/3) = -\log 3$$

$$x=2\text{の時} \quad 1\log 1/4 + 0\log(1-1/4) = \log(1/4) = -2\log 2$$

$$x=3\text{の時} \quad 0\log 5/24 + 1\log(1-5/24) = \log(19/24) = \log 19 - \log 24$$

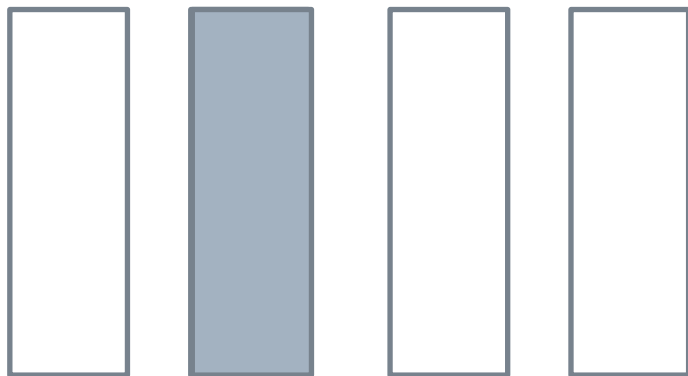
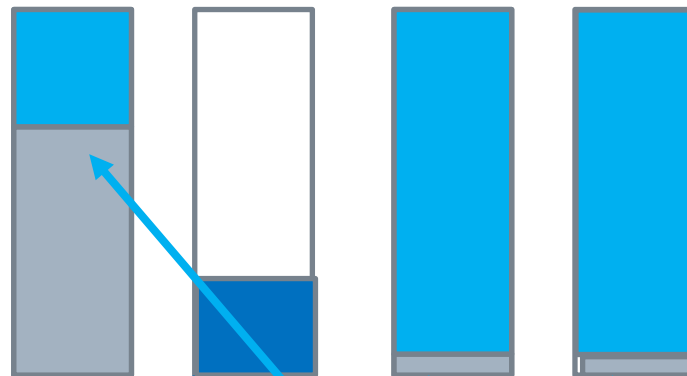
$$x=4\text{の時} \quad 0\log 5/24 + 1\log(1-5/24) = \log(19/24) = \log 19 - \log 24$$

全部足し合わせると

$$-\log 3 - 2\log 2 + \log 19 - \log 24 + \log 19 - \log 24$$

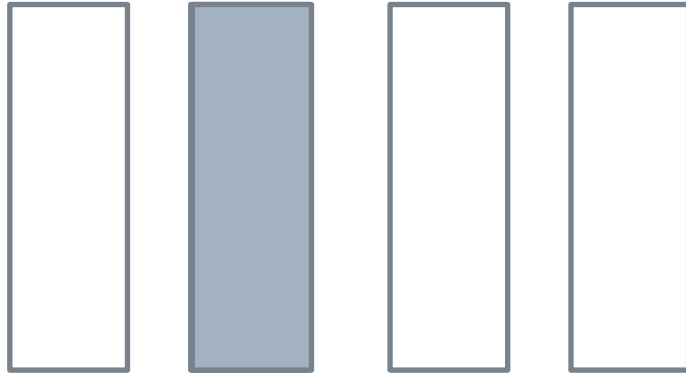
$$= -\log 3 - 2\log 2 + 2\log 19 - 2(\log 3 + \log 8) = -3\log 3 + 2\log 19 - 8\log 2$$

$$H_1(p, q) = 3\log 3 + 8\log 2 - 2\log 19$$

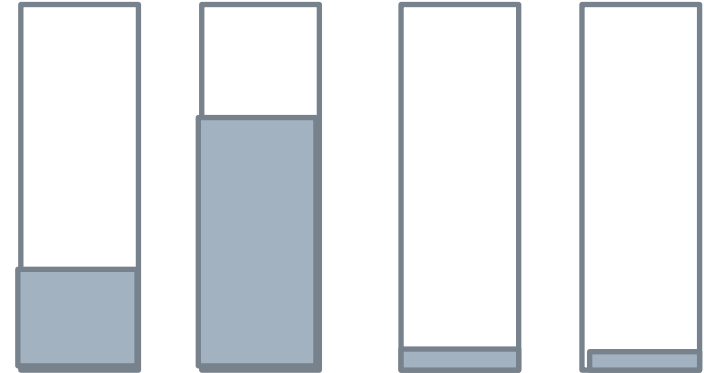
H'_1 $p(x)$ $q(x)$  $P(1)=0, p(2)=1, p(3)=0, p(4)=0$  $q(1)=2/3, q(2)=1/4, q(3)=5/24, q(4)=5/24$

$$H(p, q) = - \sum_x \{ p(x) \log(q(x)) + (1-p(x)) \log(1-q(x)) \}$$

$p(x)$ が、One Hot Valueの分布の場合でも、 $q(x)$ の分布の情報を拾っている。

H_2 $p(x)$ $q(x)$ 

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$



$$q(1)=1/4, q(2)=2/3, q(3)=5/24, q(4)=5/24$$

$$x=1 \text{ の時 } 0 \log 1/4 + 1 \log(1-1/4) = \log(3/4) = \log 3 - \log 4$$

$$x=2 \text{ の時 } 1 \log 2/3 + 0 \log(1-2/3) = \log 2/3 = \log 2 - \log 3$$

$$x=3 \text{ の時 } 0 \log 5/24 + 1 \log(1-5/24) = \log(19/24) = \log 19 - \log 24$$

$$x=4 \text{ の時 } 0 \log 5/24 + 1 \log(1-5/24) = \log(19/24) = \log 19 - \log 24$$

全部足し合わせると

$$\begin{aligned} & \log 3 - \log 4 + \log 2 - \log 3 + \log 19 - \log 24 + \log 19 - \log 24 \\ &= \log 4 + 2 \log 19 - 2(\log 3 + \log 8) = -\log 3 - 8 \log 4 + 2 \log 19 \end{aligned}$$

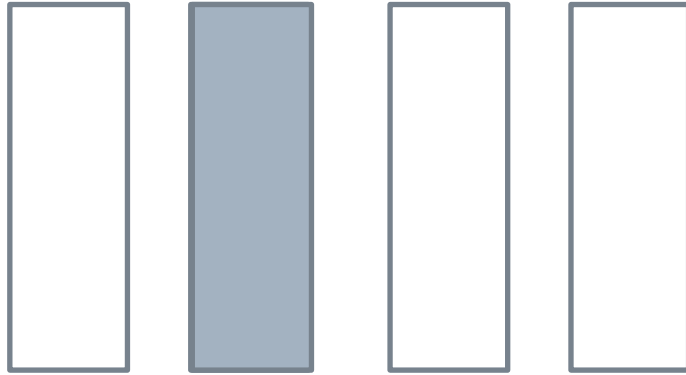
$$H_2(p, q) = \log 3 + 8 \log 2 - 2 \log 19$$

$$\begin{aligned} H_1 - H_2 &= (3 \log 3 + 8 \log 2 - 2 \log 19) - (\log 3 + 8 \log 2 - 2 \log 19) \\ &= 2 \log 3 > 0 \end{aligned}$$

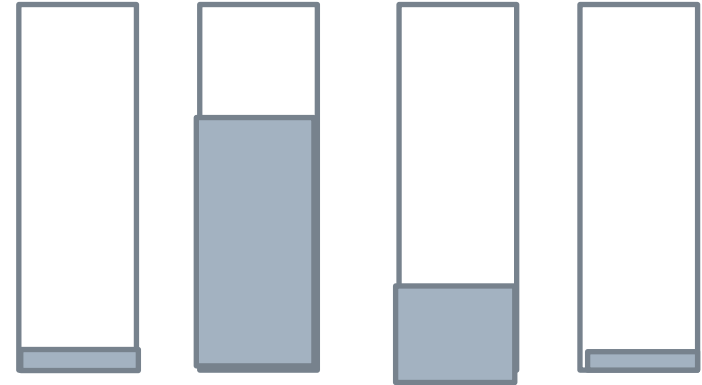
H_3

$p(x)$

$q(x)$



$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$



$$q(1)=5/24, q(2)=2/3, q(3)=1/4, q(4)=5/24$$

$$x=1 \text{ の時 } 0 \log 5/24 + 1 \log (1 - 5/24) = \log (19/24) = \log 19 - \log 24$$

$$x=2 \text{ の時 } 1 \log 2/3 + 0 \log (1 - 2/3) = \log (2/3) = \log 2 - \log 3$$

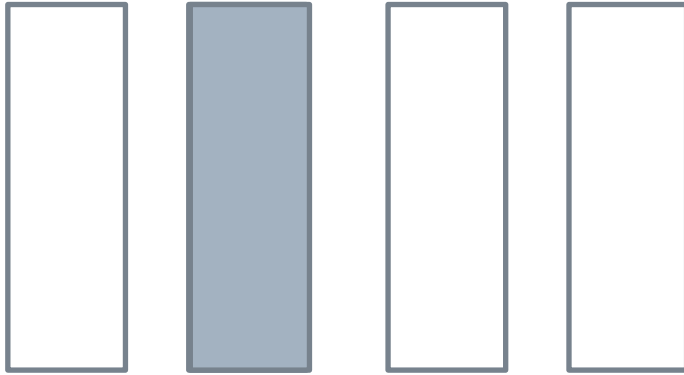
$$x=3 \text{ の時 } 0 \log 1/4 + 1 \log (1 - 1/4) = \log (3/4) = \log 3 - \log 4$$

$$x=4 \text{ の時 } 0 \log 5/24 + 1 \log (1 - 5/24) = \log (19/24) = \log 19 - \log 24$$

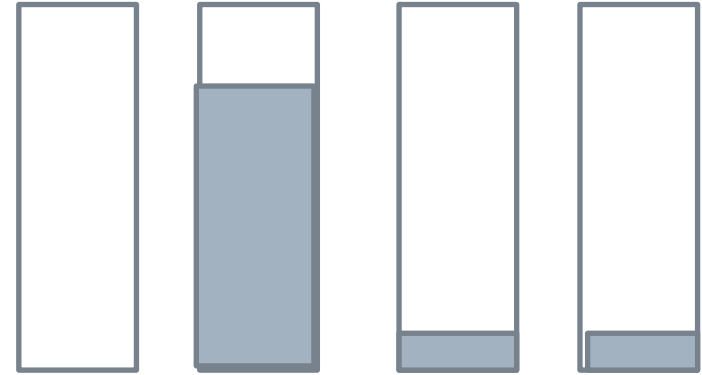
全部足し合わせると

先の H_2 と同じになる

$$H_3(p, q) = H_2(p, q) = \log 3 + 8 \log 2 - 2 \log 19$$

H_4 $p(x)$ $q(x)$ 

$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$



$$q(1)=0, q(2)=3/4, q(3)=1/8, q(4)=1/8$$

$$x=1 \text{ の時 } 0 \log 0 + 1 \log(1-0) = 0$$

$$x=2 \text{ の時 } 1 \log 3/4 + 0 \log(1-3/4) = \log(3/4) = \log 3 - \log 4$$

$$x=3 \text{ の時 } 0 \log 1/8 + 1 \log(1-1/8) = \log(7/8) = \log 7 - \log 8$$

$$x=4 \text{ の時 } 0 \log 1/8 + 1 \log(1-1/8) = \log(7/8) = \log 7 - \log 8$$

全部足し合わせると

$$\log 3 - \log 4 + \log 7 - \log 8 + \log 7 - \log 8$$

$$= \log 3 + 2 \log 7 - 8 \log 2$$

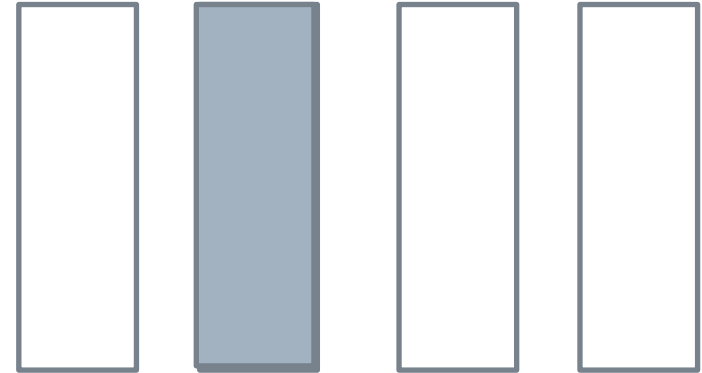
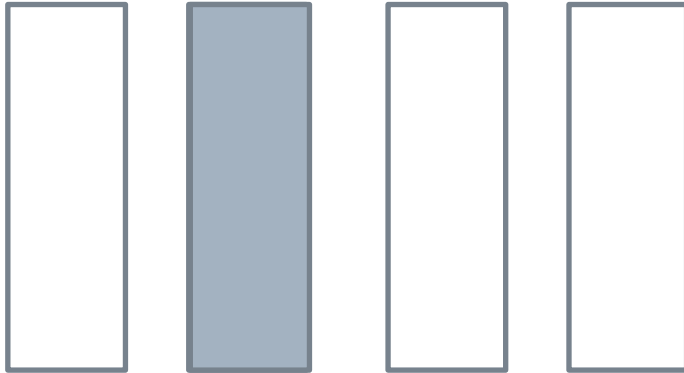
$$H_4(p, q) = 8 \log 2 - 2 \log 7 - \log 3$$

$$H_3 - H_4 = (\log 3 + 8 \log 2 - 2 \log 19) - (8 \log 2 - 2 \log 7 - \log 3) \\ = 2 \log 3 + 2 \log 7 - 2 \log 19 > 0 \quad 9 \times 49 > 19 \times 19$$

H₅

p(x)

q(x)



$$P(1)=0, p(2)=1, p(3)=0, p(4)=0$$

$$q(1)=0, q(2)=1, q(3)=0, q(4)=0$$

$$x=1\text{の時} \quad 0\log 0 + 1\log(1-0) = 0$$

$$x=2\text{の時} \quad 1\log 1 + 0\log(1-1) = 0$$

$$x=3\text{の時} \quad 0\log 0 + 1\log(1-0) = 0$$

$$x=4\text{の時} \quad 0\log 0 + 1\log(1-0) = 0$$

$$H_5(p, q) = 0$$

tensorflow/tensorflow/python/ops/nn_impl.py

https://github.com/tensorflow/tensorflow/blob/master/tensorflow/python/ops/nn_impl.py

sigmoid_cross_entropy_with_logits

$$z * -\log(\text{sigmoid}(x)) + (1 - z) * -\log(1 - \text{sigmoid}(x))$$

$$= z * -\log(1 / (1 + \exp(-x))) + (1 - z) * -\log(\exp(-x) / (1 + \exp(-x)))$$

$$= z * \log(1 + \exp(-x)) + (1 - z) * (-\log(\exp(-x)) + \log(1 + \exp(-x)))$$

$$= z * \log(1 + \exp(-x)) + (1 - z) * (x + \log(1 + \exp(-x)))$$

$$= (1 - z) * x + \log(1 + \exp(-x))$$

$$= x - x * z + \log(1 + \exp(-x))$$

For $x < 0$, to avoid overflow in $\exp(-x)$, we reformulate the above

$$x - x * z + \log(1 + \exp(-x))$$

$$\log(\exp(x)) - x * z + \log(1 + \exp(-x))$$

$$= -x * z + \log(1 + \exp(x))$$

weighted_cross_entropy_with_logits

This is like ``sigmoid_cross_entropy_with_logits()`` except that ``pos_weight``, allows one to trade off recall and precision by up- or down-weighting the cost of a positive error relative to a negative error.

The usual cross-entropy cost is defined as:

$$\text{targets} * -\log(\text{sigmoid}(\text{logits})) + \\ (1 - \text{targets}) * -\log(1 - \text{sigmoid}(\text{logits}))$$

The argument ``pos_weight`` is used as a multiplier for the positive targets:

$$\text{targets} * -\log(\text{sigmoid}(\text{logits})) * \text{pos_weight} + \\ (1 - \text{targets}) * -\log(1 - \text{sigmoid}(\text{logits}))$$

weighted_cross_entropy_with_logits

For brevity, let `x = logits`, `z = targets`, `q = pos\_weight`.

The loss is:

$$qz * -\log(\text{sigmoid}(x)) + (1 - z) * -\log(1 - \text{sigmoid}(x))$$

$$= qz * -\log(1 / (1 + \exp(-x))) + (1 - z) * -\log(\exp(-x) / (1 + \exp(-x)))$$

$$= qz * \log(1 + \exp(-x)) + (1 - z) * (-\log(\exp(-x)) + \log(1 + \exp(-x)))$$

$$= qz * \log(1 + \exp(-x)) + (1 - z) * (x + \log(1 + \exp(-x)))$$

$$= (1 - z) * x + (qz + 1 - z) * \log(1 + \exp(-x))$$

$$= (1 - z) * x + (1 + (q - 1) * z) * \log(1 + \exp(-x))$$

Setting `l = (1 + (q - 1) \* z)`, to ensure stability and avoid overflow, the implementation uses

$$(1 - z) * x + l * (\log(1 + \exp(-\text{abs}(x))) + \max(-x, 0))$$

Algorithmic Thermodynamics

John C. Baez, Mike Stay

<https://arxiv.org/pdf/1010.2067.pdf>

February 27, 2013

相対エントロピー：情報の獲得

- 有限集合 X 上の確率測度 p は、 $p \rightarrow [0, 1]$ の関数で、その和が 1 になるものである。そのエントロピーは、次の式で表される。

$$S(p) = - \sum_{x \in X} p(x) \ln p(x).$$

- しかし、他の確率測度 q に対する p のエントロピーを、次のように定義することができる。

$$S(p, q) = - \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}.$$

相対エントロピー：情報の獲得

- この「相対エントロピー」は、熱心に研究され、「Kullback-Leibler divergence」、「情報利得」を含む、多くの名前と呼ばれている。
 - 「情報利得」が、その意味をよく表している。
我々が、ある実験の結果が、確率測度 q に従って分布すると、最初には、想定していたとしよう。次に、我々が、この実験を繰り返して、この出力が測度 p に従うことを発見したとしよう。この時、取得される情報が、 $S(p, q)$ である。
-

RELATIVE ENTROPY IN BIOLOGICAL SYSTEMS

JOHN C. BAEZ AND BLAKE S. POLLARD

<https://arxiv.org/pdf/1512.02742.pdf>

2016/02/12

Information Geometry (Part 6)

John Baez

http://math.ucr.edu/home/baez/information/information_geometry_6.html

2011/01/21

-
- I like relative entropy because it's related to the [Bayesian interpretation of probability](#). The idea here is that you can't really 'observe' probabilities as frequencies of events, except in some unattainable limit where you repeat an experiment over and over infinitely many times. Instead, you start with some hypothesis about how likely things are: a probability distribution called the [prior](#). Then you update this using [Bayes' rule](#) when you gain new information. The updated probability distribution — your new improved hypothesis — is called the [posterior](#).
 - ~~□ And if you don't do the updating right, you need a swift kick in the posterior!~~

-
- For example, suppose you're flipping a coin, so your set of events is just $\Omega = \{\text{heads}, \text{tails}\}$. In this case all the integrals are just sums with two terms.

Suppose your **prior** assumption is that the coin is fair. Then $q(\text{heads}) = 1/2$, $q(\text{tails}) = 1/2$.

But then suppose someone you trust comes up and says "Sorry, that's a trick coin: it always comes up heads!" So you update our probability distribution and get this **posterior**: $p(\text{heads}) = 1$, $p(\text{tails}) = 0$.

$$S(p, q) = 1 \cdot \log\left(\frac{1}{1/2}\right) + 0 \cdot \log\left(\frac{0}{1/2}\right) = 1$$

So: you've learned *one bit of information!*

-
- That's supposed to make perfect sense. On the other hand, the reverse scenario takes a bit more thought.
 - You start out feeling sure that the coin always lands heads up. Then someone you trust says “No, that’s a perfectly fair coin.” If you work out the amount of information you learned this time, you’ll see it’s *infinite*.

prior: $q(\text{head})=1, q(\text{tail})=0$

posterior: $p(\text{head})=1/2, p(\text{tail})=1/2$

$$S(p, q) = \frac{1}{2} \cdot \log((1/2)/1) + \frac{1}{2} \cdot \log((1/2)/\mathbf{0}) = \infty$$

-
- why is that?

-
- The reason is that something that you thought was impossible — the coin landing tails up — turned out to be possible. In this game, it counts as infinitely shocking to learn something like that, so the information gain is infinite. If you hadn't been so darn sure of yourself — if you had just believed that the coin *almost always* landed heads up — your information gain would be large but finite.
 - The Bayesian philosophy is built into the concept of information gain, because information gain depends on two things: the prior and the posterior. And that's just as it should be: *you can only say how much you learned if you know what you believed beforehand!*

"The category of probabilistic mappings -- With Applications to Stochastic Processes, Statistics, and Pattern Recognition"

V. W. Lawvere

<http://goo.gl/53nxEv>

1962年

Stochastic Processes and Decision Maps

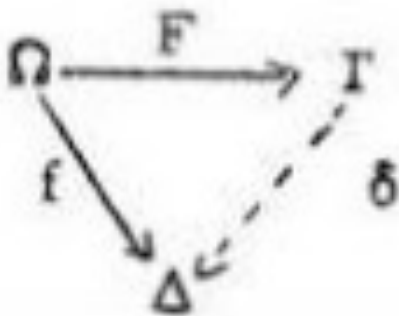
「確率過程と決定写像」

- Ω をある空間とする。 Δ を、 Ω を「パターン」ごとに分割した空間とする。
 - 「パターン認識」とは、 Ω から Δ の写像 f を見つけること。
 - ただし、一般には、 Ω は巨大で、全体をスキャンするのも、 f を見つけるのも困難である。
 - そこで、実際に「観測された状態」からなる Γ を考える。
 - さらに、 Ω から Γ への、次のような確率写像 F を考える。
 Ω の基本的な状態の一つを ω とする時、 Γ の部分集合 A に対して、 $F(\omega, A)$ は、 A についての条件付きの確率を与える。
-

Stochastic Processes and Decision Maps

「確率過程と決定写像」

- δ を、 Γ から Δ の写像とする。
この時、「パターン認識」の問題は、完備化されれば、 $f = F \cdot \delta$ となるような、最良の写像 δ を求める問題に帰着される。
- こうした δ を求める問題は(中略)、ある量を最大化する最適化問題として定式化できる。(彼は、こうした「量」を、具体的に提示している。)



参考文献

- Brendan Fong: "Causal Theories: A Categorical Perspective on Bayesian Networks"
<http://arxiv.org/pdf/1301.6201v1.pdf>
 - John Baez, Brendan Fong, Tobias Fritz, Tom Leinster: "Network Theory III: Bayesian Networks, Information and Entropy" <http://goo.gl/wXG8uQ>
 - John C. Baez, Tobias Fritz, and Tom Leinster: "A Characterization of Entropy in Terms of Information Loss" <http://goo.gl/p13C4d>
-