
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

この間、「embeddingの共有・蓄積・交換の世界の拡大」という共通の視点から、

- マルレク「機械の言語能力の獲得を考える」
- マルレク「embeddingプログラミングの基礎」

という二つのセミナーを開催してきました。

マルレク「「知識のハブ」の変化を考える」は、これら二つのセミナーに続く、embedding シリーズの第三弾です。

Agenda

「知識のハブ」の変化を考える

Part 1　「知識のハブ」とは何か

Part 2　グローバルな検索の世界の変化

Part 3　MaruLabo KnowledgeHub の紹介



Part 1

「知識のハブ」とは何か

Agenda
「知識のハブ」とは何か

- はじめに
- 情報ネットワークの新しい構造
- なぜ、「知識のハブの変化」を考えるのか？
- 「知識」のどのような領域に注目するのか



はじめに

「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

はじめに

セッション「情報ネットワークの新しい構造」

「情報ネットワークの新しい構造」のセッションは、SNS等えの「コミュニケーションのハブ」や、そのより強い表現である「ハブとしてのメディア」を侵食する、新しい構造がネットワークの世界で生まれているのではないかという問題意識を述べた試論です。

基本的には、「言語能力・意味を理解する能力」を持った機械の登場が、コミュニケーション・チャンネルにどのような変化を引き起こすのかを考えようということなのですが、まだ、説得力のあるものには展開できていないことは自覚しています。

ここで直接扱っているのは、コミュニケーション・チャンネルの技術論です。ただ、そこに基礎を持つ従来型の「コミュニケーション論」「メディア論」は、「言語能力を持つ機械」の台頭という、現実の非連続的な巨大な変化の中で、その有効性・妥当性を急速に失うだろうと考えています。

ここでは、 Shannonの通信チャンネルモデル を「参照枠」として利用しています。

セッション「なぜ、知識のハブの変化」を考えるのか？」

「なぜ、「知識のハブの変化」を考えるのか？」のセッションでは、次のような議論が展開されます。

この一時的なメッセージと永続するメッセージというメッセージの性質の違いは、一般的な「コミュニケーション・ハブ」と今回のセミナーで取り上げる「知識のハブ」の違いを特徴づけるものです。

セッション「「知識」のどのような領域に注目するのか」

「「知識」のどのような領域に注目するのか」のセッションでは、次のような議論が展開されます。

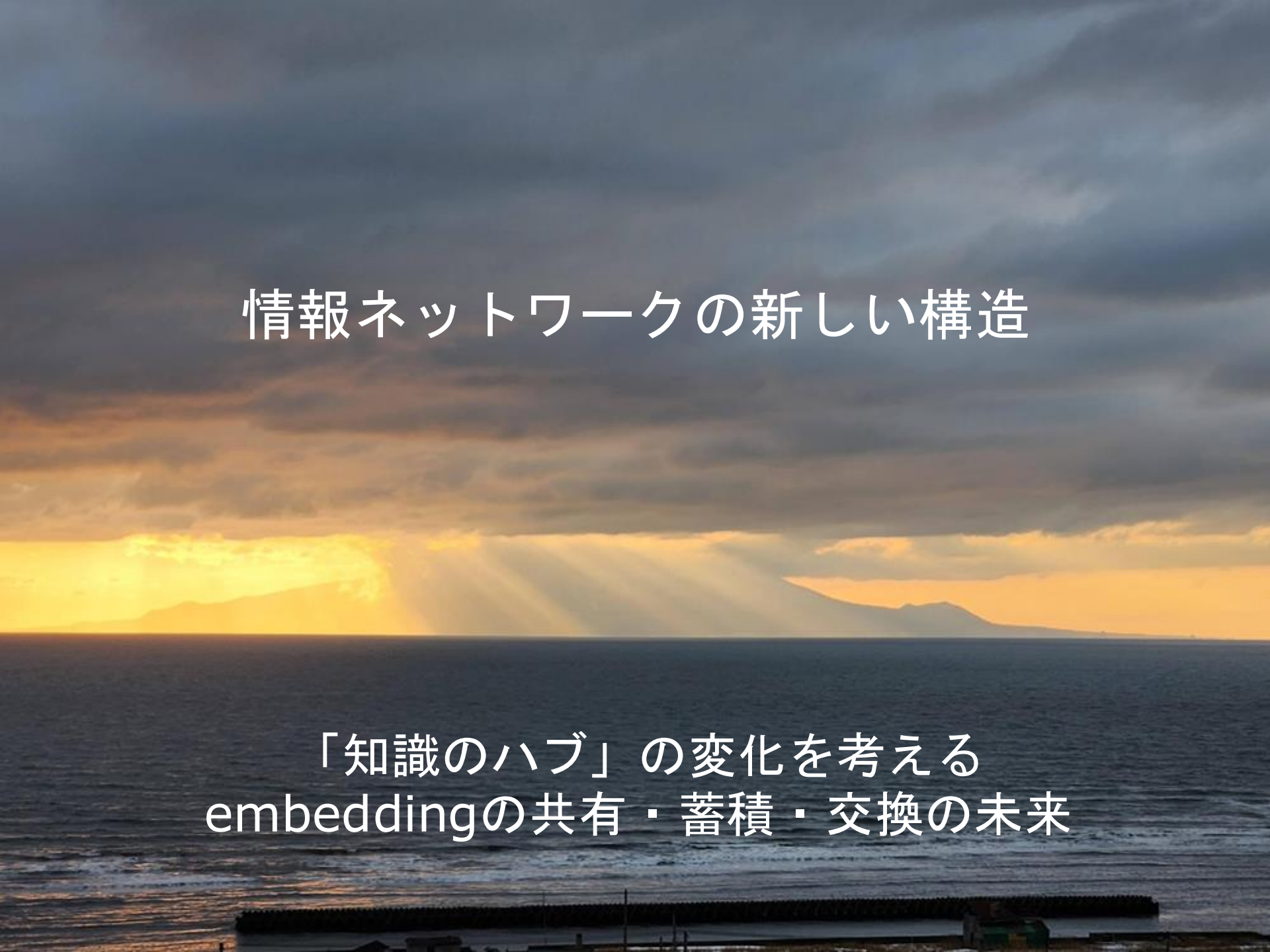
このセッションでは、二つの領域に注目します。

第一の領域は、人間中心だった知識の世界に「言語能力を持った機械」が登場したことによって大きな影響を受けている領域です。

この領域は、情報の共有・蓄積・交換の社会的様式とそれについての社会的意識の急速な変化によって特徴づけられます。

第二の領域は、「言語能力を持った機械」の登場とは相対的に独立に、知識の世界での人間と機械の境界面の新しい拡大が見られる領域です。

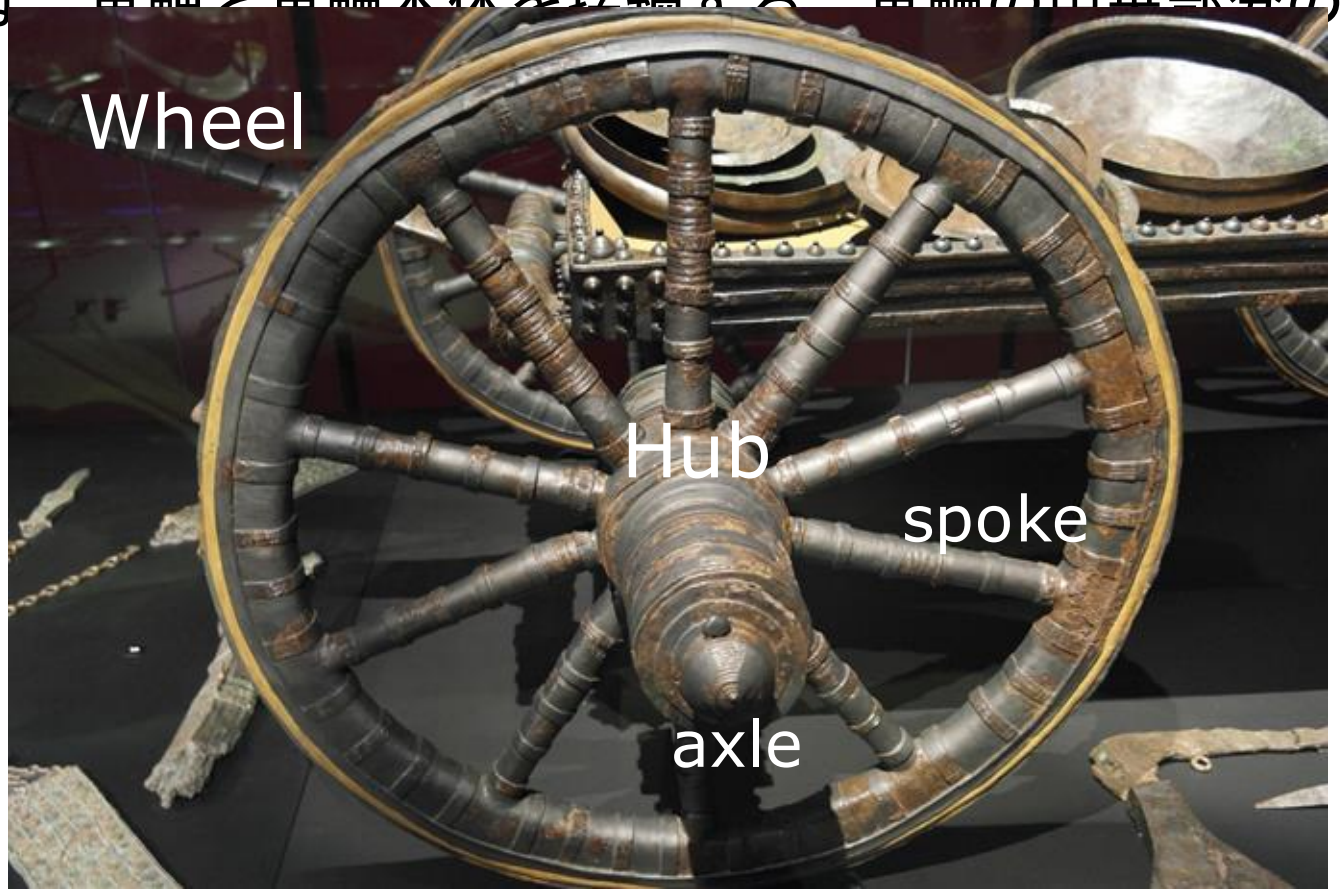
この領域は、人間の自然認識や数学的認識といった「累積的知識」の中核的領域を含んでいます。


情報ネットワークの新しい構造
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

ネットワークと「ハブ」の働き

様々な「ハブ」

「ハブ」というのは、元々は、「車輪」と結びついた言葉です。
ハブは、車軸と車輪本体を接続する、車輪の中央部分のことです。











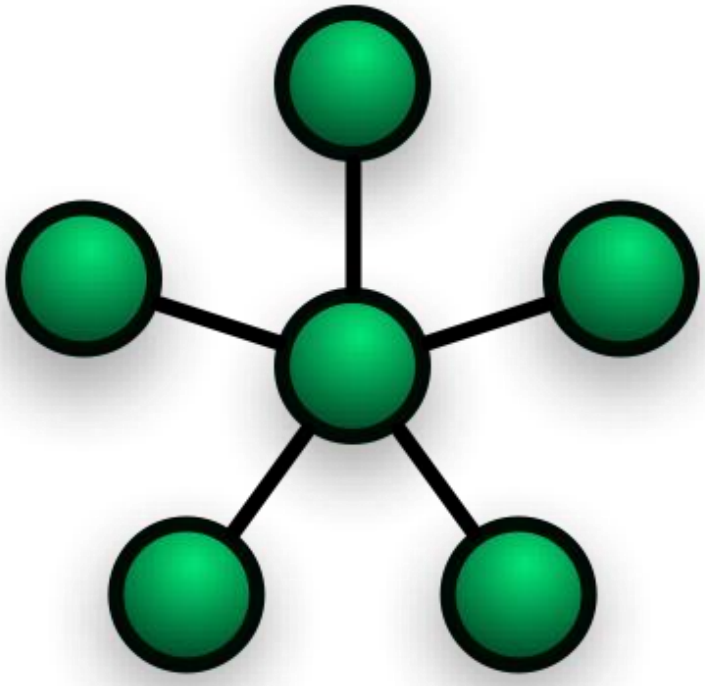






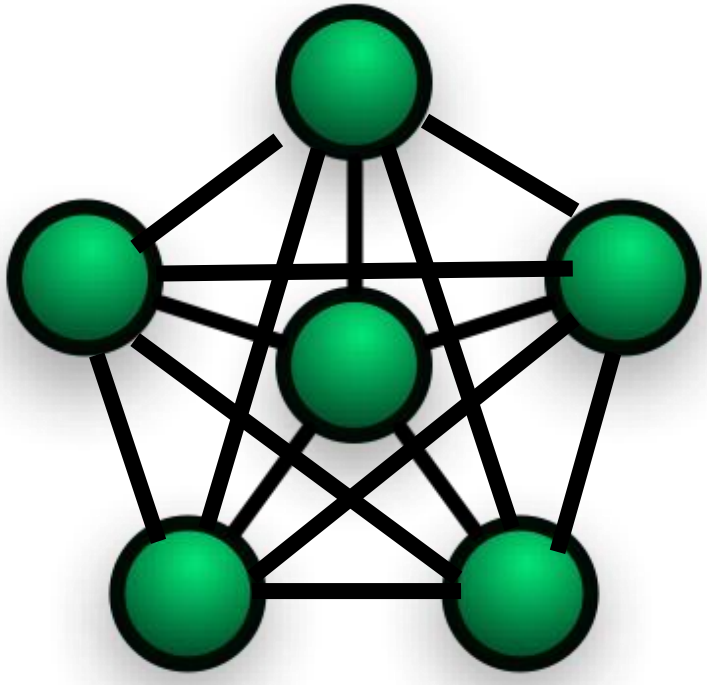
https://en.wikipedia.org/wiki/Wheel より

Hub and Spoke Topology



中央のハブと、スポークの先に繋がったそれぞれのノードが結ばれるネットワーク・トポロジー

PtoP Topology



任意のノードと任意のノードが結ばれるネットワーク・トポロジー

我々に身近なHub構造

- 交通のハブ
  - Airline hub
  - 公共交通
- 物流のハブ
  - 海上輸送
  - Cargo airline
- 産業のハブ
  - Seatle(Boeing)
  - Silicon Valley
  - 豊田(TOYOTA)
- 放送
  - テレビ
  - ラジオ
- Contents Delivery Network
  - YouTube, Netflix, ...
- GitHub

- イーサーネット hub
- USB hub

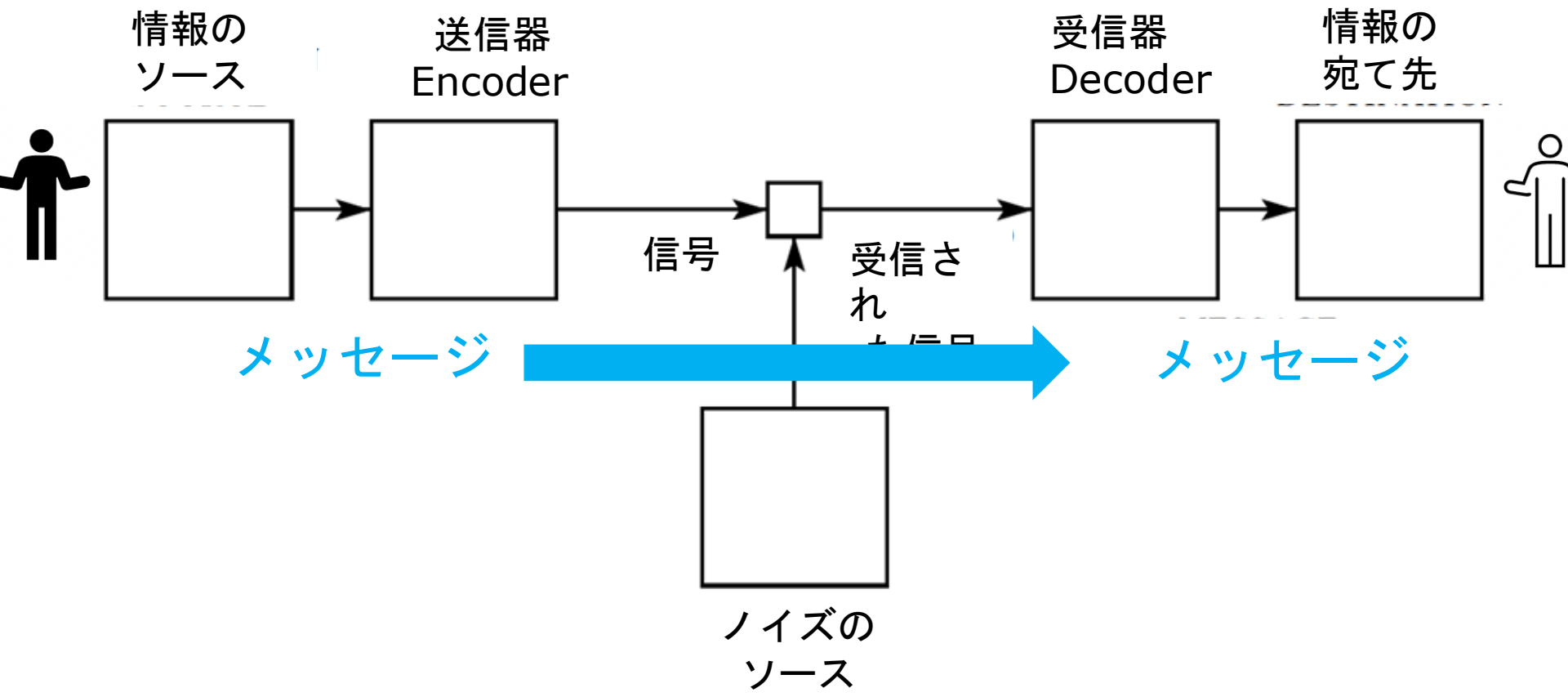
いずれのモデルでも基礎にあるのは…



いずれのモデルでも、その基礎にあるのは、ノードとノードの一対一の関係である

情報ネットワークの構造
Shannonの通信チャンネルモデル

情報ネットワークの構造
Shannonの通信チャンネルモデル



通信の基本的な問題は、情報のソースで選択されたメッセージを、正確であれ近似的であれ、情報の宛て先で再生産することである。

メッセージは意味を持っている
しかし、「意味」は通信システムでは重要ではない

多くの場合、そのメッセージは意味を持っている。すなわち、そのメッセージは、あるシステムに従って確かな物理的実体あるいは概念的実体を、参照するかそれに関連している。

ただ、こうしたの意味論的な諸側面は、工学的問題とは無関係である。重要な通信的側面は、実際のメッセージは可能なメッセージの集合の中から選択された一つのメッセージだと言うことである。

通信システムは、実際に選択されたメッセージだけに対してではなく、全ての可能な選択に対して機能するように設計されていなければならない。なぜなら、設計の時点では、どのメッセージが選ばれるかは、わからないからである。

A Mathematical Theory of Communication　C. E. SHANNON

http://www.qsl.net/n9zia/pdf/shannon1948.pdf　1948年

なぜ、「意味」は通信システムでは重要ではないのか
もう一つの理由

先のShannonの言葉には、もう一つの理由がある。当時は、あまりにも自明なことなので、当然の前提としてあえて触れていないだけだと思う。

それは、意味のあるメッセージを生み出すことが出来るのも、そのメッセージの意味を解釈することが出来るのも、人間だけであるということである。そうした能力を持った人間は、通信システムの外部にいる。

もちろん、そうした「意味を扱う」人間の「言語能力」は、通信システムには帰属しない。

Shannonの通信路モデルと
安全な「コミュニケーションのハブ」への応用

このShannon由来の通信路モデルは、現代では意識されることは少ないのですが、あらゆる情報のディジタル化の基礎になりました。それは、現在のネットワーク上の「コミュニケーションのハブ」と呼ぶべきもの、例えば、ほとんどすべてのSNSに応用されています。

同時に、この通信路モデルは、その誕生時から符号化の理論を通じて、暗号理論と深く結びついていました。それは、効率的な情報圧縮の限界を示し、あわせて、情報を隠すための安全な通信の限界を明確に定義しました。

Shannonの理論は、通信システムの全ての世代交代を貫いて、その上での安全で高速な通信の基礎を提供しています。

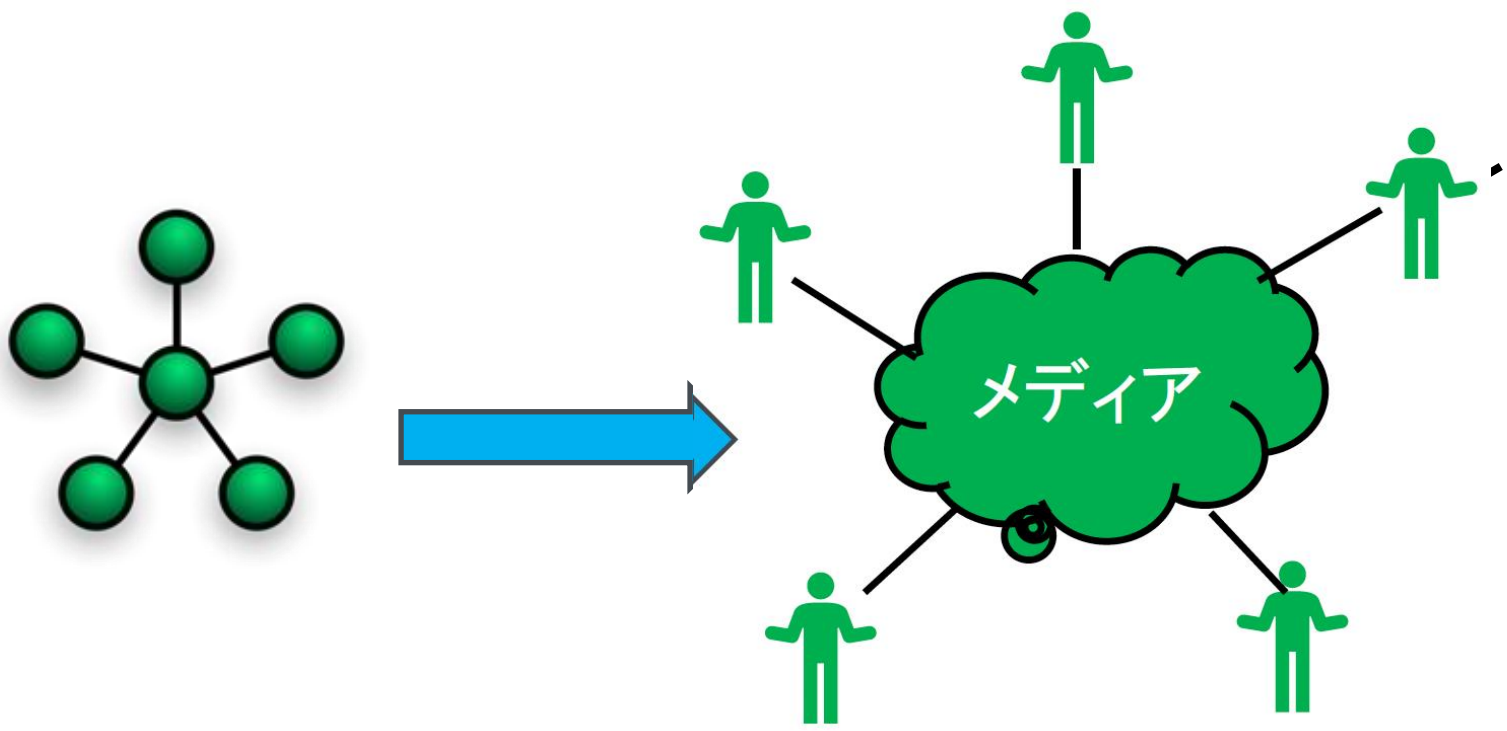
Shannonの通信路モデルと
情報理論的エントロピー概念の成立

重要なことは、Shannonが通信システムの実践的な分析を通じて、情報量＝エントロピーという新しい理論的概念に到達したことです。

ある意味、驚くべきことですが、20世紀電話の通信品質の改善という技術的取り組みの中で発見されたこの情報理論的エントロピー概念は、蒸気機関の熱効率の改善という技術的努力の中で発見された、19世紀の熱力学的エントロピー概念と繋がるものでした。

エントロピー概念の拡大とその各分野への浸透は、現代の科学技術の大きな特徴となっています。

「ハブとしてのメディア」

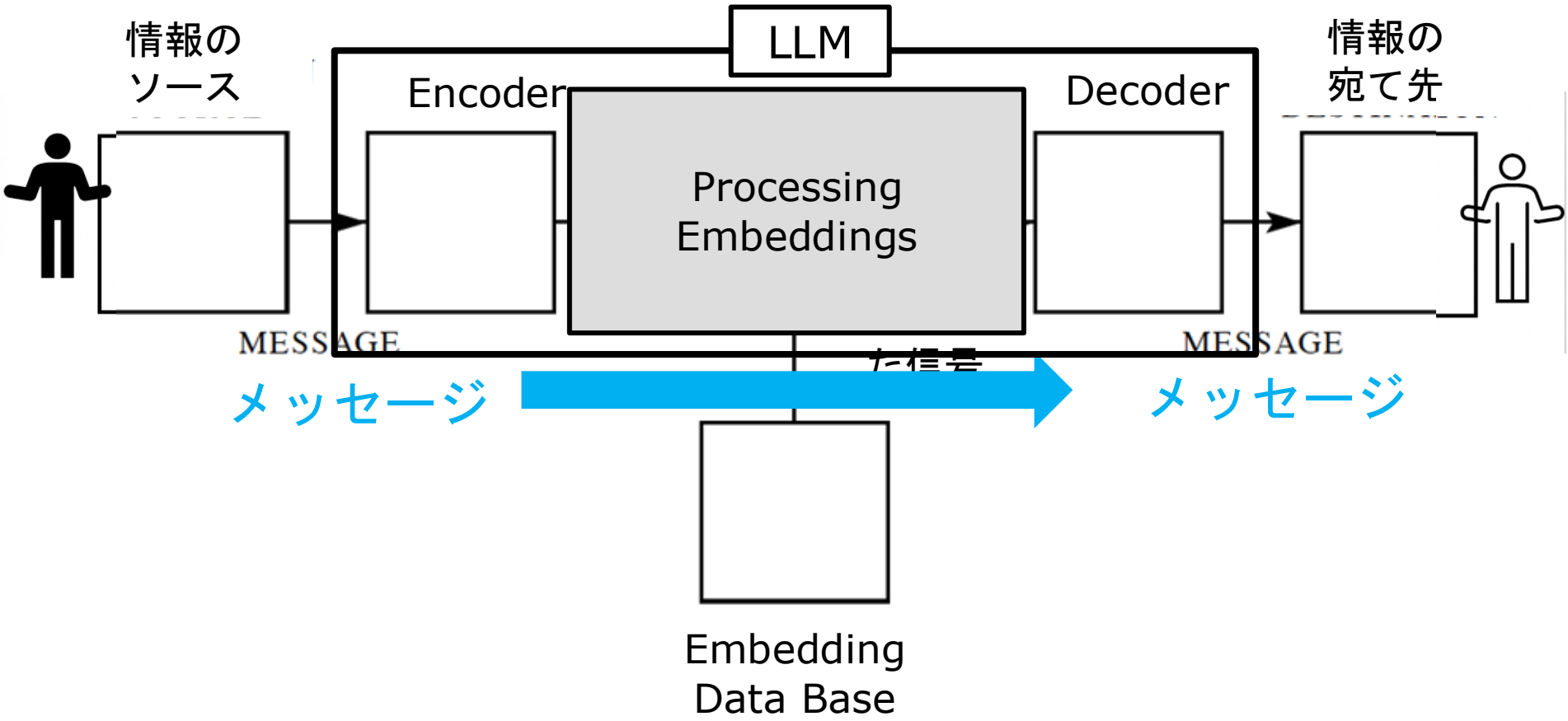


ハブ

としてのメディア

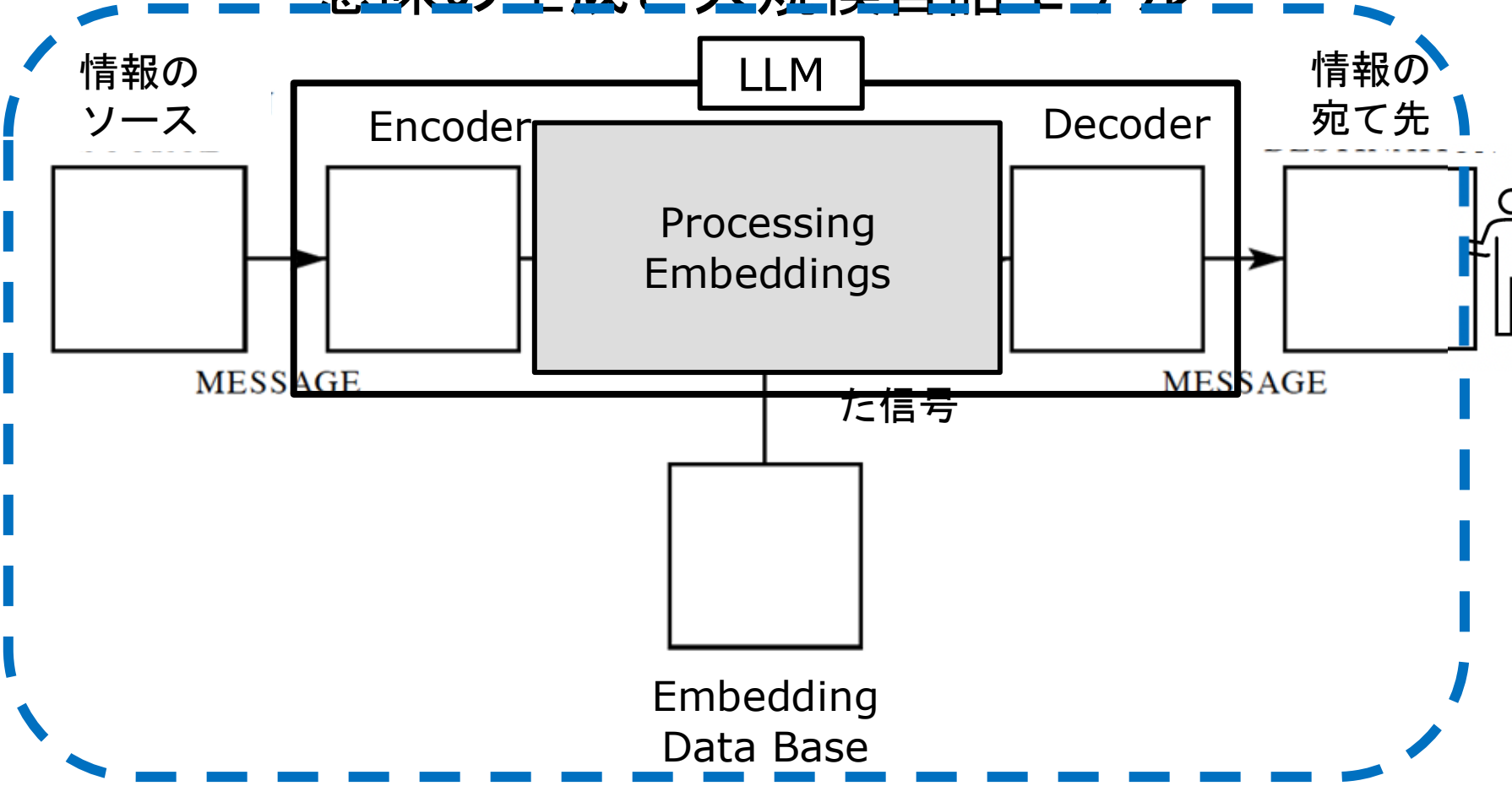
情報ネットワークの新しい構造
意味の生成と大規模言語モデル

情報ネットワークの新しい構造
意味の生成と大規模言語モデル



このネットワークの課題は、ソースの情報の宛て先での再生産ではなく、新しい意味を持ったメッセージを生産することである。

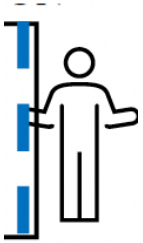
情報ネットワークの新しい構造
意味の生成と大規模言語モデル



このネットワークの課題は、ソースの情報の宛て先での再生産ではなく、新しい意味を持ったメッセージを生産することである。

情報ネットワークの新しい構造
意味の生成と大規模言語モデル





このネットワークの課題は、ソースの情報の宛て先での再生産ではなく、新しい意味を持ったメッセージを生産することである。

情報ネットワークの新しい構造
意味の生成と大規模言語モデル

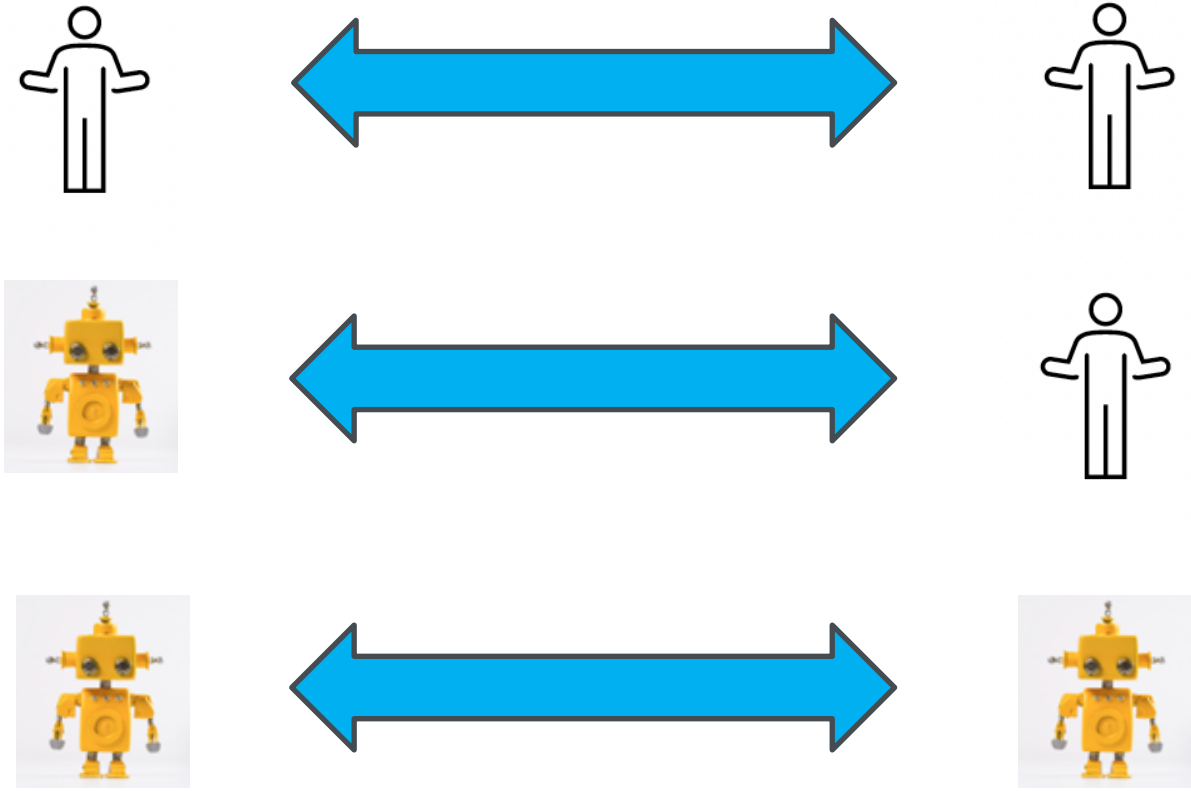


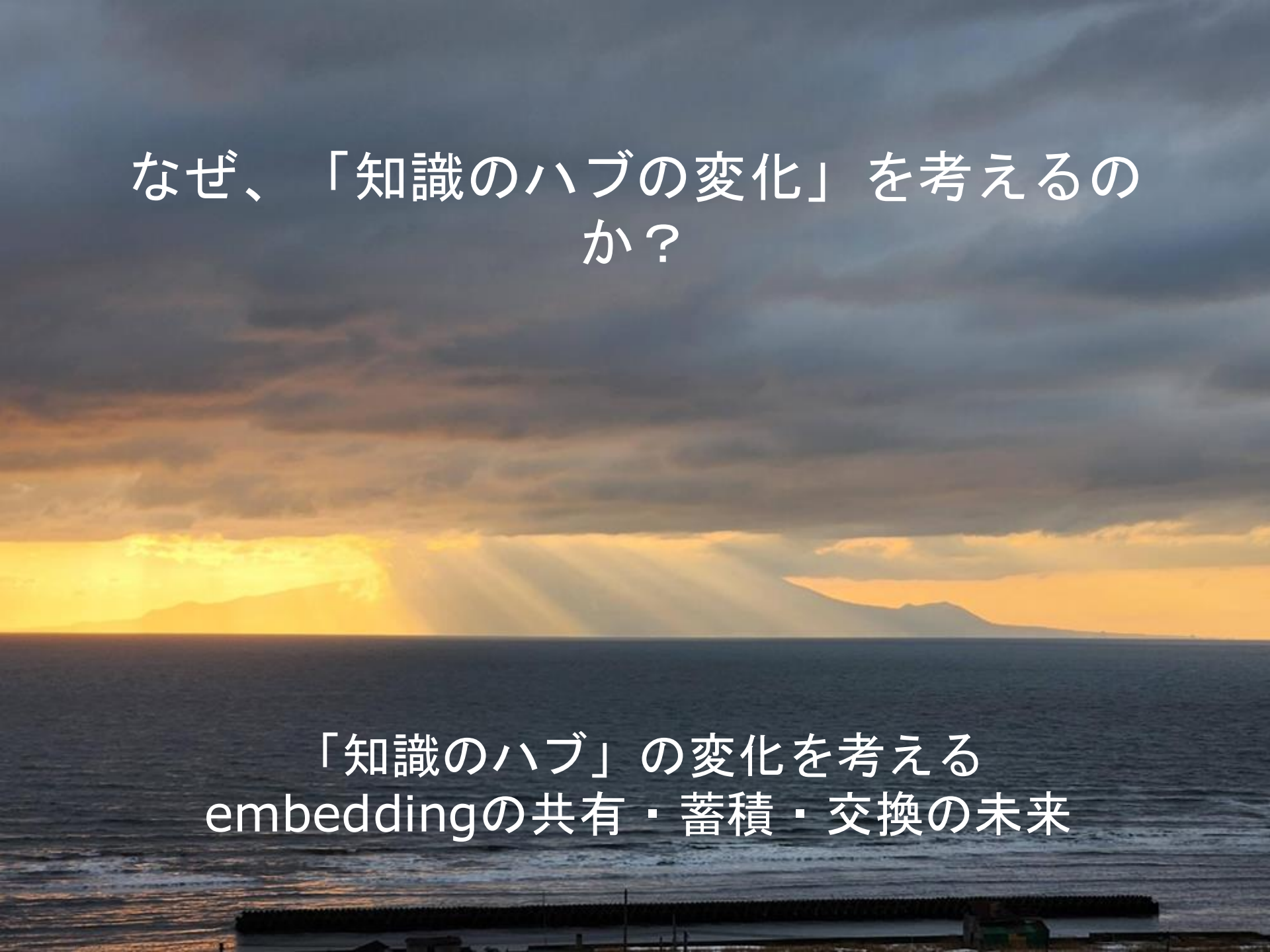
メッセージ



このネットワークの課題は、ソースの情報の宛て先での再生産ではなく、新しい意味を持ったメッセージを生産することである。

メッセージ交換のいくつかのパターン




なぜ、「知識のハブの変化」を考えるのか？
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

なぜ、「知識のハブ」の変化を考えるのか？

僕が、 「機械の言語能力の獲得」という巨大な変化の中で、特に「知識のハブ」の変化に注目するのには、大きく言って二つの理由があります。

第一に、SNSのような人間と人間の日常的な「コミュニケーション・ハブ」や、Webや伝統的な「メディア・ハブ」の世界は、生成AI技術の台頭の中でも、基本的には、その構造を維持していると、僕は考えているからです。現在注目すべき変化は、そこで生まれているわけではありません。

確かに、Fake NewsやFake Image/Video と言った重要な問題は存在します。それは、別途考えていきたいと思います。

一時的なメッセージと永続するメッセージ「コミュニケーション・ハブ」と「知識のハブ」 の違い

僕が「知識のハブ」の変化に注目する第二の理由は、embedding技術の拡大を中核とする現在進行中の変化が、temporalな永続性を持たない一時的なメッセージ主体の日常的なコミュニケーションに与える影響は少ないにしても、科学や技術といった累積的な性質を持つ人間の知識の体系に与える影響は、大きいと考えているからです。

この一時的なメッセージと永続するメッセージというメッセージの性質の違いは、一般的な「コミュニケーション・ハブ」と今回のセミナーで取り上げる「知識のハブ」の違いを特徴づけるものです。

知識の世界への機械の登場

一時的なメッセージを扱うコミュニケーションの理論で、永続的なメッセージを扱う知識の理論を構築できないと僕は考えています。

また、「文字の発明」「印刷技術の登場」「学校・教育システムの整備」「インターネットの登場と普及」等々といった、エポックメイキングな「メディア」の登場・成立史を中心とする従来の「メディア論」に、単純には包摂されない変化が、ネットワーク上で起きつつあると僕は考えています。

なぜなら、注目すべき基本的な変化は、人間しかいなかった「知識」の世界に。言語能力を持った機械が登場しようとしていることだからです。

AIと「意識の変化」についての二つのエピソード

このセミナーでは、「知識のハブ」の「変化」について考えます。ただ、具体的な議論に入る前に、AIがもたらす変化とそれを受容する人間の意識の変化について、二つの特徴的なエピソードを紹介しようと思います。

一つは、人間はAIについて正反対の意見を、それが矛盾していると意識することなく受け入れることができるというエピソードです。

ここでは、これまでも何度か紹介したことのある、「機械は考えることができるか」という問題についての人間の意識の変化についてのTuringの素晴らしい予言を取り上げます。

このエピソードは、人間が矛盾した考えを受け入れるようになるには、長い時間が必要だという平凡なまとめが可能かもしれません。

「長い時間が必要？」
それでは、意識の変化は、いつ訪れるのでしょう？

今回紹介する、もう一つのエピソードは、AIについての意識の変化は突然訪れるということについてのエピソードです。

2011年夏に公刊されたIEEE Intelligent Systems誌の「AIの殿堂」 記事のエピソードが特に興味深いのは、AIの専門家にとってさえ予期できないスピードで、 AIについての意識の変化が訪れることがあるということです。

1950年　Turingの予言

人間はAIについて正反対の意見を、
それが矛盾していると意識することなく
受け入れることができる



機械は考える事が出来るか？

「私は、『機械は考える事が出来るか？』という問題を考察することを提案する。」

今から75年以上前、膨大な数の真空管からなるコンピューターがこの世に生をうけてまもなく、チューリングは、先の一節ではじまる「計算機械と知能 」と題する、現代の人工知能（Artificial Intelligence） 研究の第一ページをしるす、興味深い論文を雑誌「マインド」に寄稿します。

“COMPUTING MACHINERY AND INTELLIGENCE”
http://www.loebner.net/Prizef/TuringArticle.htm

Turingの考察の特徴

この 「機械の思考」の可能性についての、チューリングの考察をふりかえって見ると、それは、予想される反論への反論という奇妙な構成をしていることに気づきます。

そのことは、彼がこの論文で主要に念頭に置いていたのが、「機械は、そもそも考えることなどできない」 といった意見のバリアントであった事をはっきりと示しています。

予想される反対論 の最も重要な例

彼が、予想される反対論の筆頭に、思考は神が人間にのみ与えた不滅の魂の働きでありそうしたものを与えられていない機械や動物は考える事ができないという 形而上学的な「神学的反対論」をあげたのは偶然ではないように恩います。

「こうした（考える）機械をつくろうと試みるにしても、我々は魂を創造する（create）神の力を不敬にも奪っている事にはならない。それは、子供を出産(procreation)したからといって神の力を奪っているとはいえないのと同じである。むしろ、我々は、いずれの場合も神の創造する魂にすみかを提供する神の意思の道具なのである。しかし、こうした事は全くの思弁にすぎない。」

彼自身が反対論を組み立て、彼自らが反論する。この執拗な繰りかえしは、ユーモアの衣をまとってはいるのですが、我々に、彼のある種のラジカリズムを印象づけるのです。

そこでは、問題は、むしろ世界観的レベルに属するといっても過言ではない「信念」の問題として語られています。彼は、「機会の思考は可能だ」という「信念」を語っています。

「約50年のうちには、……普通の質問者が5分間質問したあとで正解率が70%を超えない様にコンピュターをプログラムすることが可能であると私は信じている。」

ここで、本当は、この論文の重要な内容である Turing Test のことを触れないといけないのですが、今回はその部分についての考察は割愛します。

「信念」と「科学的探究」

この論文で、「信念」と「科学的探究」についてTuringが述べている見解は、非常に興味深いものです。

「私は、こうした信念を秘密にする事が、いかなる有用な目的にも役立つものではないとさえ信じている。

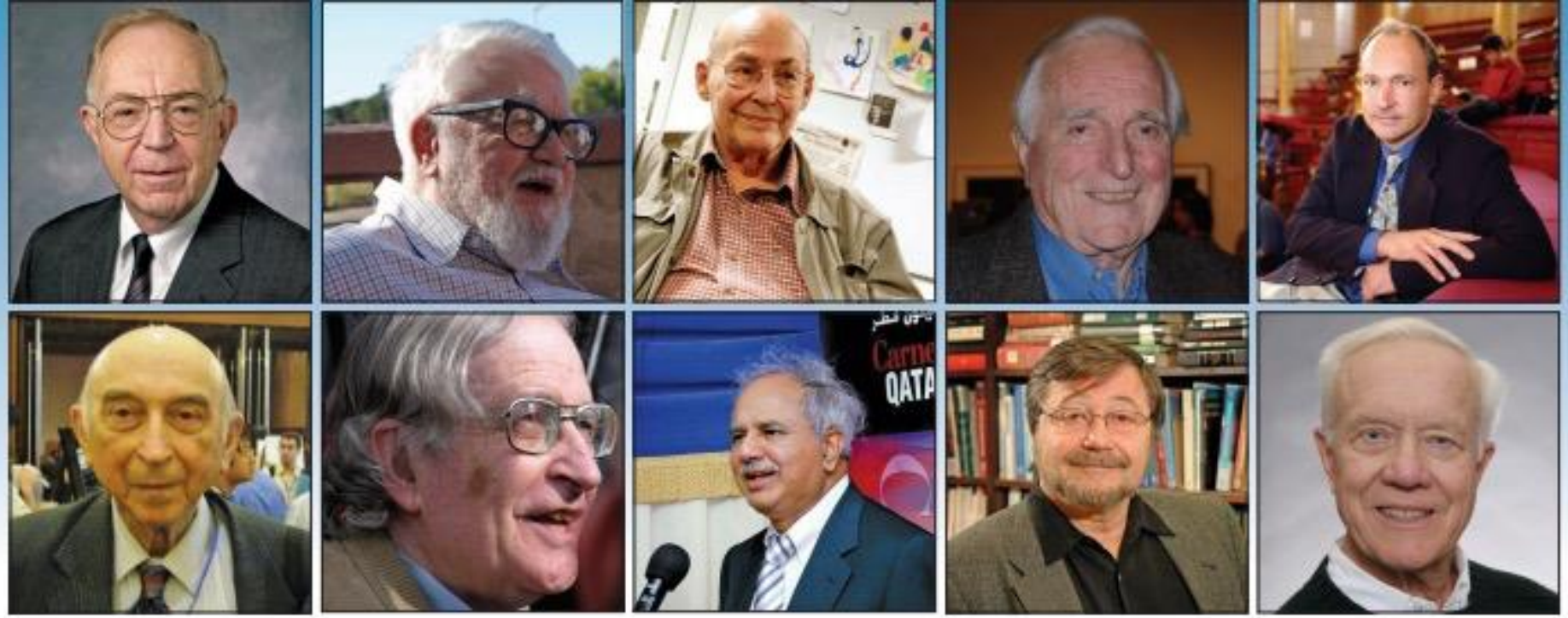
科学者は、証明されていない、いかなる予想の影響をも決して受け入れることなく、はつきりと確立され予想から、はっきりと確立された事実へと冷厳に歩みをすすめるという俗見は、全く誤っている。

様々な予想は、それらが探求の有用な方向を示唆する故に、非常に重要なものである。」

Turingの予言

「機械は考える事が出来るか？」という、最初に掲げた間題が、今では議論にも値しない程無意味なものである事は私も認めよう。

しかし、それにもかかわらず、今世紀の終わりには人達の意見が大きく変化して、ついには、矛盾していると考えることなく『機械の恩考』について語ることができるようになるであろうと私は信じている。



我々は、変化に気づかないことがあるということ
2011年のIEEEの「AIの殿堂」の事例

AI's Hall of Fame

2011年 IEEE Intelligent Systems 誌 特集
AI’s Hall of Fame （AIの殿堂）

「2010年、創刊25周年を記念して、IEEE Intelligent Systemsの編集委員会および諮問委員会は、AIおよびインテリジェントシステム分野に多大な貢献を果たした先駆者たちへの感謝と敬意を表し、同分野および社会に与えた顕著な影響と功績を称えるため、「IEEE Intelligent Systems 殿堂」を設立することを決定いたしました。本特集では、IEEE Intelligent Systems 殿堂の初代殿堂入りを紹介いたします。」

- **Edward Feigenbaum**
  Engineering Knowledge
- **John McCarthy**
  Logic and Common Sense
- **Marvin Minsky**
  Cognitive Science and AI
- **Douglas Engelbart**
  AI and Interactive Computing
- **Tim Berners-Lee**
  Collective Knowledge
- **Lotfi Zadeh**
  Fuzzy Logic and Computational Intelligence
- **Noam Chomsky**
  Computational Linguistics and Cognitive Science
- **Raj Reddy**
  AI and Societal Impact
- **Judea Pearl**
  Probability, Causality, and Intelligence
- **Nils J. Nilsson**
  Problem Solving and Planning

2012年に起きたこと

2012年に起きたこと

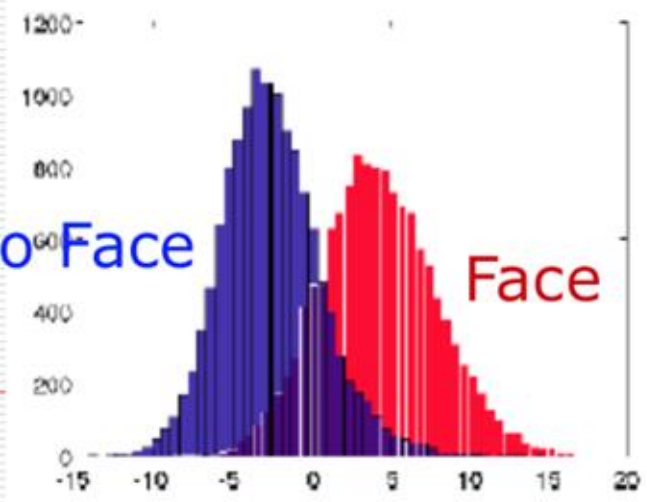
現在の人工知能に対する関心の高まりは、2012年に始まります。この年、

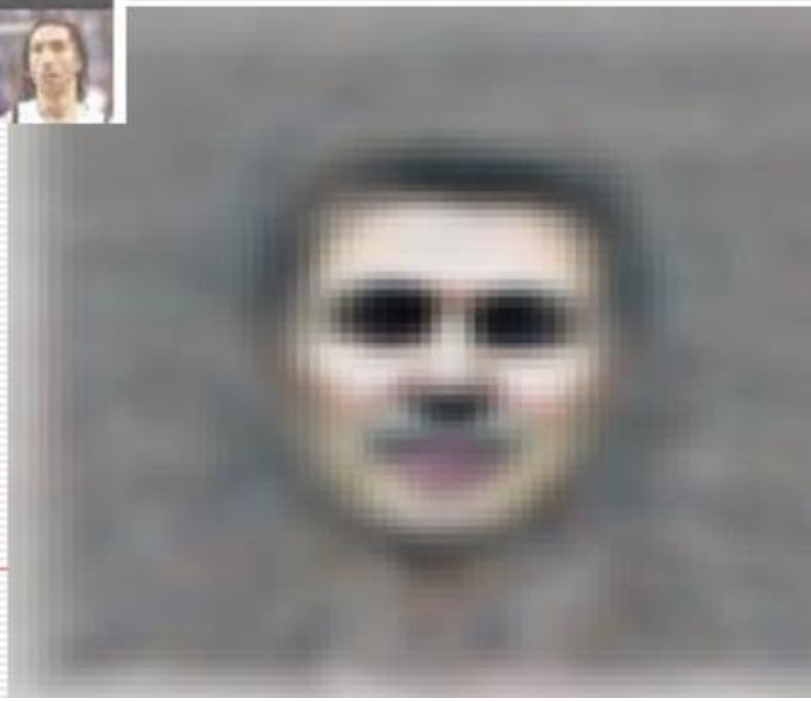
1. Googleの16,000コアの巨大なシステム DistBeliefが、教師なしでネコと人間の顔を学習するのに成功した。
2. また、ImageNetの画像認識コンテストで、Convolutional Neural Network (CNN）のAlexNetが、圧勝した。
3. さらに、Tront大+Microsoft+Google+IBM 四者の共同研究チームが、音声認識にディープラーニングを用いた手法を公開。従来の方式を、はるかに上回る精度を達成した。

こうして、2012年を重要なターニング・ポイントとしてニューラル・ネットワーク上のディープラーニングという手法は、現代のAI技術の中核の座を占めるに至なりました。

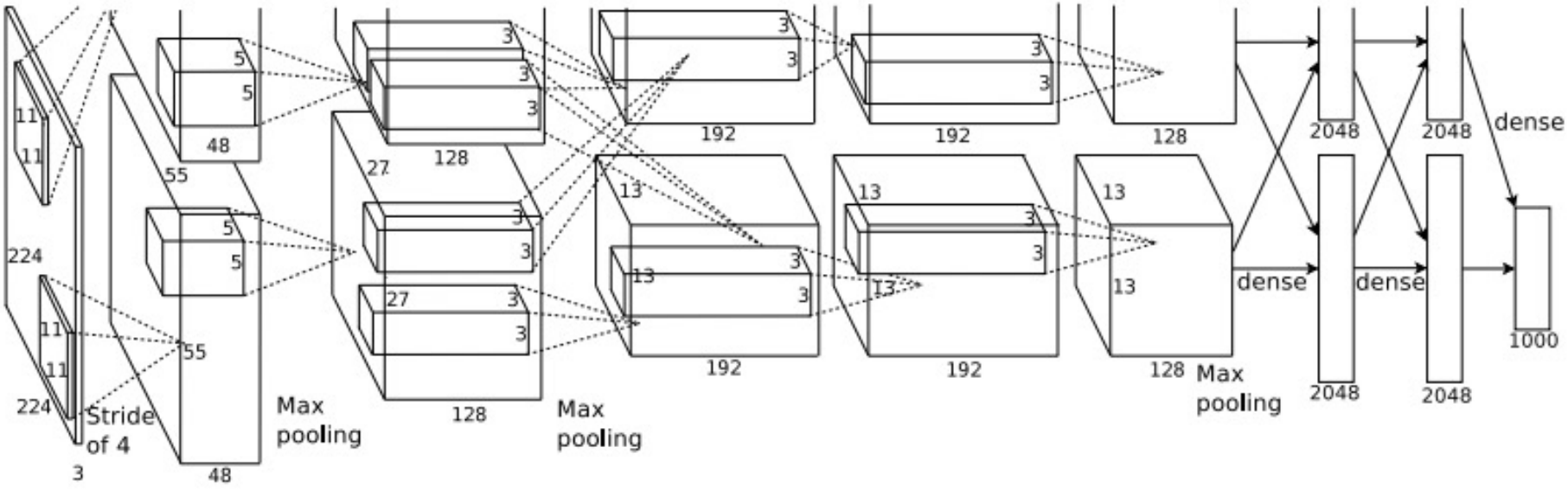


これが、一千万個の画像データから、16,000コアのマシンが三日間かけて抽出した「人間の顔」のイメージらしい。子供の絵に似ているのが面白い。





AlexNet


224
224
3
11
11
Stride
of 4
55
55
48
48
5
5
Max
pooling
27
27
128
128
3
Max
pooling
13
13
192
192
192
192
128
128
Max
pooling
2048
2048
2048
2048
dense
dense
dense
1000

“Deep Neural Networks for Acoustic Modeling in Speech Recognition”
Geoffrey Hinton et al.

ディープラーニングを用いた
音声認識技術の飛躍

https://goo.gl/GT6Bg9

2012年







2012年 Deep Learningの爆発的登場



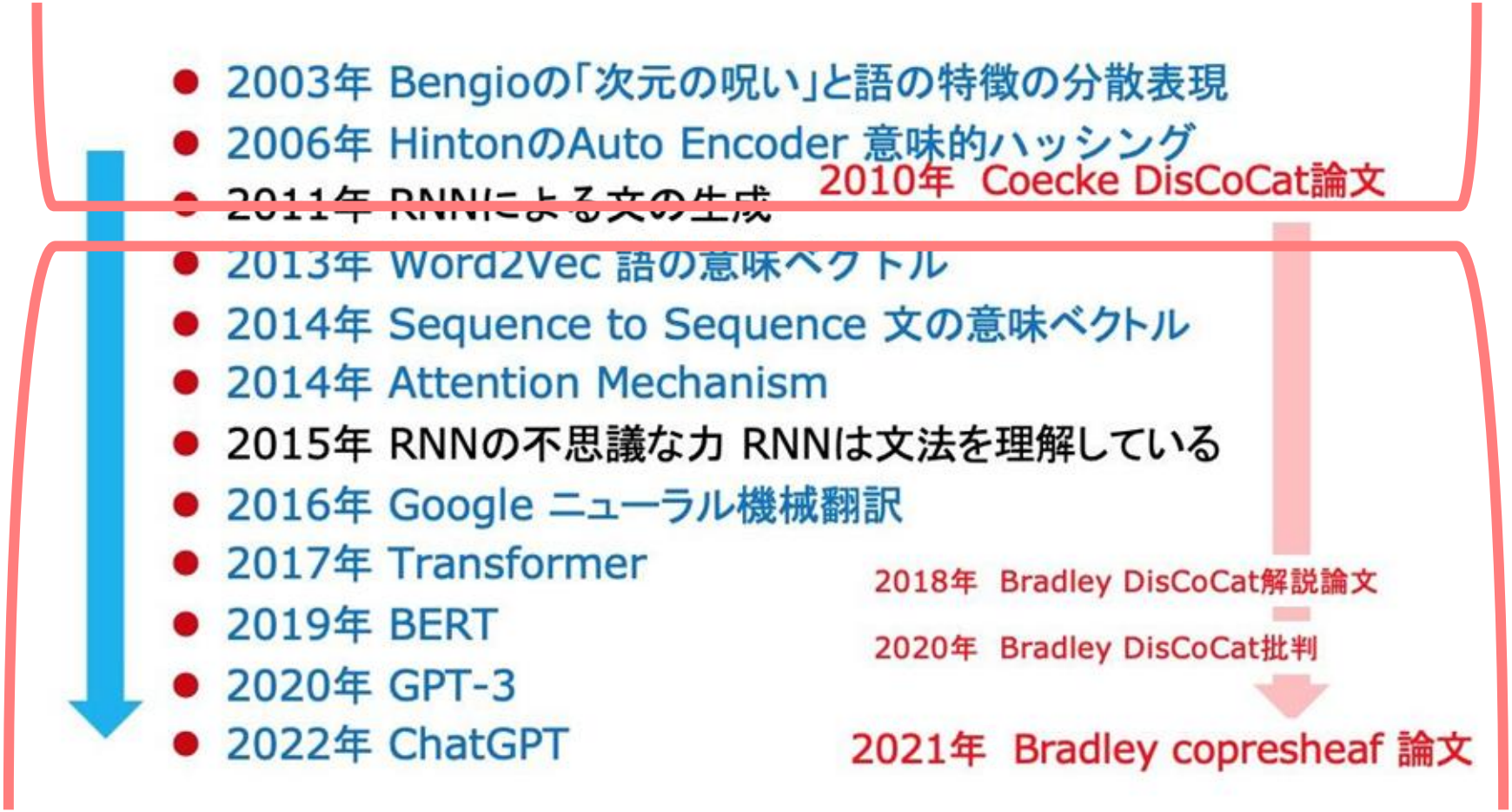
Geoffrey Hinton
Yann LeCun
Yoshua Bengio

こうした成功にもかかわらず、ConvNetは、2012年のImageNetコンテストまで、コンピューター・ビジョンと機械学習のコミュニティーから、ほとんど無視されていた。

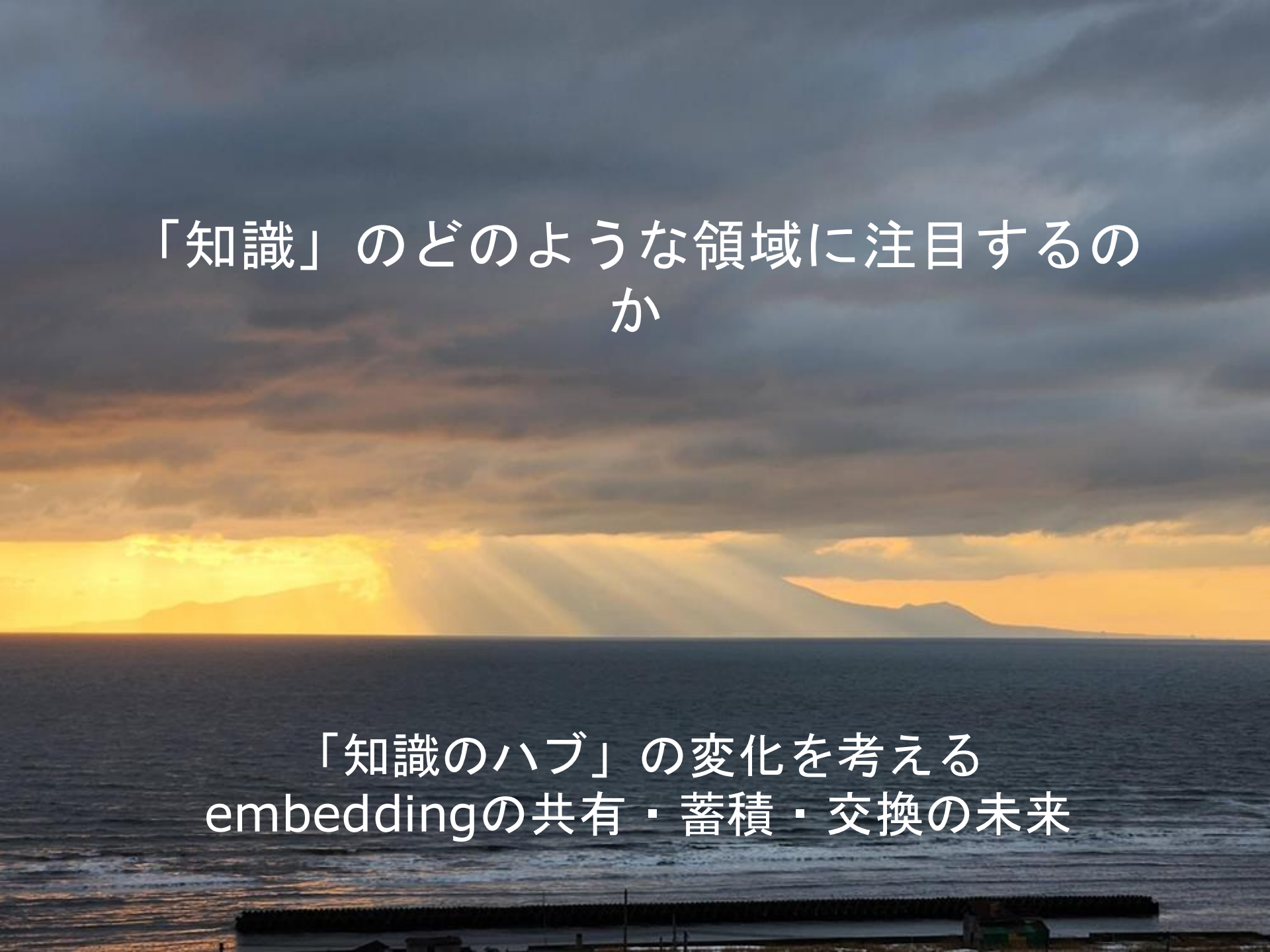
この時、Webからとられた、1,000の異なったクラスを含んだ約100万の画像のデータセットに、ディープ・コンボリューショナル・ネットワークが適用され、驚嘆すべき結果を残した。最高の競争相手のアプローチの、ほとんど半分のエラー率を達成したのだ。

2015年 “Deep Learning” Hinton, LeCun, Bengio
https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf

2011年 IEEE 「AIの殿堂」
2012年 Deep Learning 大爆発



このリストは、大規模言語モデルの成立にフォーカスしたもの


「知識」のどのような領域に注目するのか
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

「知識」のどのような領域に注目するのか

このセッションでは、「知識のハブ」の変化を考えるうえで、どのような「領域」に注目するのかを考えてみようと思います。

このセッションでは、二つの領域に注目します。

今回のセミナーが注目する二つの領域

第一の領域は、人間中心だった知識の世界に「言語能力を持った機械」が登場したことによって大きな影響を受けている領域です。

この領域は、情報の共有・蓄積・交換の社会的様式とそれについての社会的意識の急速な変化によって特徴づけられます。

第二の領域は、「言語能力を持った機械」の登場とは相対的に独立に、知識の世界での人間と機械の境界面の新しい拡大が見られる領域です。

この領域は、人間の自然認識や数学的認識といった「累積的知識」の中核的領域を含んでいます。

今回のセミナーのアプローチ
「変化」の特徴的なエピソードを集める

今回のセミナー「「知識のハブ」の変化を考える」では、様々な領域に現れている「変化」の特徴的なエピソードを紹介したいと考えています。

それは、異なる「場所」に起きた、バラバラなエピソードの羅列に見えるかもしれません。

それが、理論的整理のための試論としての、「変化を考える」という、今回のセミナーがとったアプローチです。

ただ、恣意的でランダムに見え、それらのエピソードの選択は、大きくいうと、先の「今回のセミナーが注目する二つの領域」に属しています。

- 「言語能力を持つ機械」の登場によって社会的インパクトを受ける領域
- 自然認識・数学的認識の発展における人間と機械の境界面が変化している領域

以下で、今回のセミナーで、この二つの領域のエピソードとして、どのようなトピックを取り上げるのか、その概要を述べてみたいと思います。

詳細は、後続のセッションで展開します。

「言語能力を持つ機械」の登場による社会的インパクト

ここでは、次のようなトピックを取り上げます。

- グローバルな検索

ネットワーク上のほとんどすべての情報は、人間だけでなく言語能力を持つ機械にとっても利用可能な形で共有・蓄積され、多くの人間は、「検索」を通じて、機械と人間が共有する情報にアクセスするようになります。

それは、一面では専門的な知識の「民主化」と捉えることもできます。

- 学会と論文

かつては、知識の蓄積・交換の中心だった「学会」組織も、その権威への静かな挑戦に直面しています。直接的には、AIが生成した論文の急増、また、査読プロセスへのAIの利用を巡る議論の拡大等々。

僕は、論文の作成にAIを利用することを排除する必要はないと思っ

- 教育と学習

より深い問題は、制度としての教育組織と個人としての学習の、それぞれの制約と可能性の問題です。個人としての学習の比重が大きく増大する方向に、歴史的なシフトが起きるだろうと考えています。

ただ、現実問題としては、現在のような過渡期に、将来にわたる展望を語ることに難しさを感じています。

- Big AI TechとAI研究者たち

「知識のハブ」として、Big AI Techを取り上げることに、違和感を覚える人は少なくないと思います。ただ、現在の最も影響力のある「知識のハブ」の一つは、Big AI Techだと、僕は考えています。そこでは、知識の「交換」は、AI研究者の「移動」によって担われているように見えます。

その限りでは、それは安定したシステムではありません。

自然認識・数学的認識の発展における人間と機械の境界面の変化

ここでは、次のようなトピックを取り上げます。

- **巨大観測装置**

「知識のハブ」として、CERNのAtlas実験やJames Webb Space Telescopeのような巨大観測装置を取り上げることに首をかしげる人もいると思います。しかし、機械が今日のような言語能力を持つ以前から、科学者のネットワークのいくつかは、こうした巨大な観測装置を中心に組織されていました。

近年の注目すべき動向を紹介したいと思います。

- **数学と形式的証明**

僕は、近年注目を一身にあびている「機械の言語能力の獲得」以外にも、機械にはもう一つの重要な能力、「数学的推論能力」を持っていると考えています。重要なことはそれは、「言語能力」には還元されない能力です。






Part 2
グローバルな検索の世界の変化

Agenda
グローバルな検索の世界の変化

- AIと検索
- ベクトル検索とキーワード検索とのハイブリッド検索
- Sparse embedding



グローバルな検索の世界の変化

AIと検索

「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

グローバルな検索の世界の変化

このパートでは、AIの登場がグローバルなインターネット上の検索に与える影響を見ていきたいと思います。

検索の世界は、今、急速な変化の只中にあります。その変化を特徴づけるのは、機械が「ことばの意味を理解し始めた」という変化です。その変化は、「文字から意味への変化」と考えることができます。

「機械と人間が意味を共有するembeddingの世界」の登場が、検索の世界を変えようとしています。

AIと検索

人間が持つほとんどすべての情報は、ネットワーク上に集約されようとしています。

重要なことは、人間だけでなく言語能力を持つ機械であるAIにとっても利用可能な形で、こうしたネットワーク上の情報は、共有・蓄積されていると言うことです。

多くの人間は、「検索」を通じて、機械と人間が共有する情報にアクセスするようになります。

ネットワーク上の情報検索の能力は、人間にとって最も身近な機械＝Aiの能力として現れます。

Google Co-Founder Larry Page Predicts the Future of Search With AI (2000)



https://youtu.be/1OT_Uj2z3Z0

2000年のラリー・ページの予想

「人工知能は、いわば究極のGoogleのようなものになるでしょう。

つまり、究極の検索エンジンが実現すれば、ウェブ上のあらゆる情報を理解し、ユーザーが何を求めているかを正確に把握して、適切な情報を提供してくれるようになるのです。

これは明らかに人工知能の役割であり、基本的にどんな質問にも答えられるようになるでしょう。というのも、ほぼすべての情報がウェブ上に存在するからです。

とはいえ、現時点ではその実現には程遠い状況ですが、少しずつその目標に近づいていくことは可能です。そして、それがまさに私たちが取り組んでいることです。」


グローバルな検索の世界の変化
ベクトル検索とキーワード検索との
ハイブリッド検索
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

前回のセミナーのふりかえり

前回のセミナー「embeddingプログラミングの基礎」
https://www.marulabo.net/docs/embedding2/　の
Part 3 「embeddingと検索技術の新しい展開」の

- 「文字列検索と画像検索」
  https://youtu.be/cMlAlVRv44w?list=PLQIrJ0f9gMcMtd68N6t0r-R-V_keMWFhH

- 「Vector Search 」
  https://youtu.be/Ks1NkPEAjHU?list=PLQIrJ0f9gMcMtd68N6t0r-R-V_keMWFhH

をご覧ください。

Googleの検索ふりかえり

Googleの検索は、Googleのcrawlerがインターネット上で網羅的に収集した文字列の巨大な集まりAから、文字列xを含むAの要素のリストを作ることから始まります。

それは、ある言葉が本の何ページ目に出てくるかをまとめる、本の末尾の索引（Index）ページを作る作業に似ています。

ただ、Googleの検索の働きは逆で、ある言葉が与えられた時、その言葉が出てくるページを探します。同じIndexづくりでも、それを「逆引きIndex」ということがあります。

Googleの検索は、「逆引きIndex」で作成された、ある文字列xを含む文字列のリストに優先度をつける「ページ・ランク」アルゴリズムとそれを計算する大規模分散システム「MapReduce」によって支えられていました。

加えて、その巨大な文字列検索システムは「広告モデル」という、ビジネスモデルと結びついていました。

Webスケールに対応した「文字列の検索」技術は、21世紀初頭のIT技術の進化を牽引した技術です。

問題は、この「文字列検索」から「意味検索」への変化が、今まさに進行中だということです。

ベクトル検索ふりかえり
ANN: Approximate Nearest Neighbor

ベクトル検索は、データ(テキスト、画像、動画)を連続的な多次元空間内のベクトル – embedding として表現します。

「検索」とは、クエリベクトルに幾何学的に最も近い近傍を見つける問題(Nearest Neighbor Search)になります。

数十億のベクトルの中から厳密な最近傍を見つけることは、線形探索では、計算コストが高すぎます。

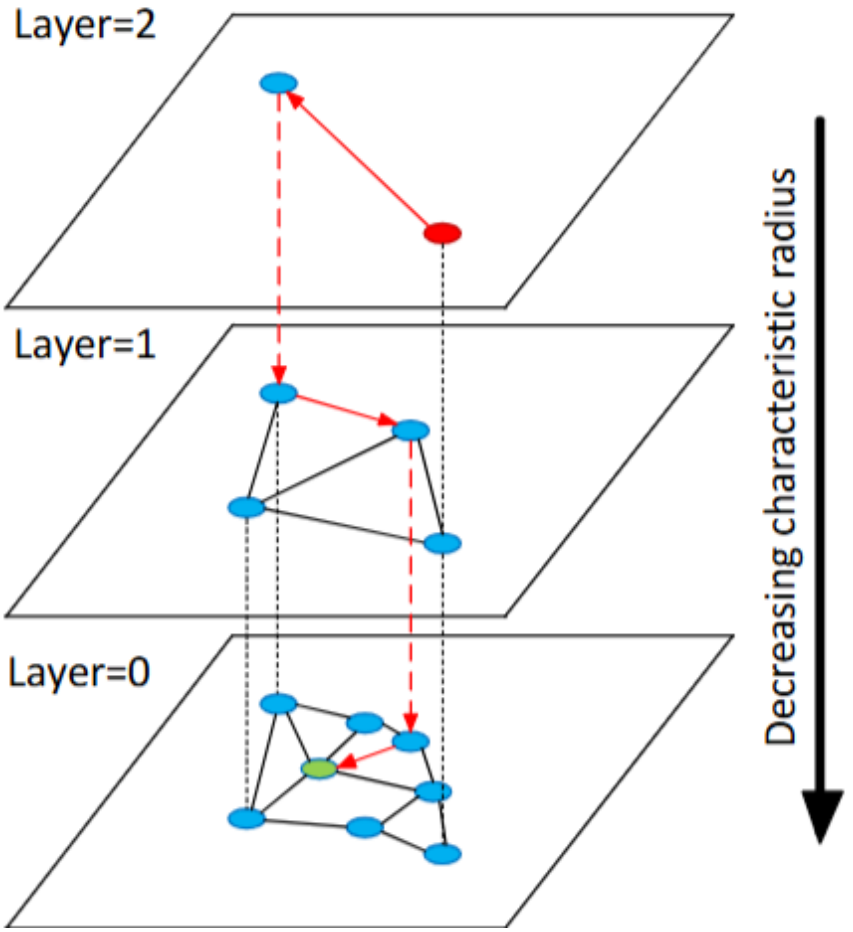
そのため、現代のベクトル検索システムは近似最近傍探索(ANN: Approximate Nearest Neighbor)アルゴリズムを使用しています。

近似最近傍探索(ANN)と HNSW (Hierarchical Navigable Small World)

HNSW (Hierarchical Navigable Small World)は、 現在のベクトルインデックスの業界標準です。

これは多層的なグラフ構造を作成し、「スモールワールド」ネットワークの特性を利用して、検索は最上層(粗いビュー)から始まり、下層(細かいビュー)へとズームインしていくことで、対数時間での検索を可能にしています。

階層型NSWアルゴリズムの概念図



検索は最上位層の要素（赤色で表示）から開始される。

赤い矢印は、エントリポイントからクエリ（緑色で表示）へのGreeding アルゴリズムの進行方向を示している。

Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs
https://arxiv.org/abs/1603.09320

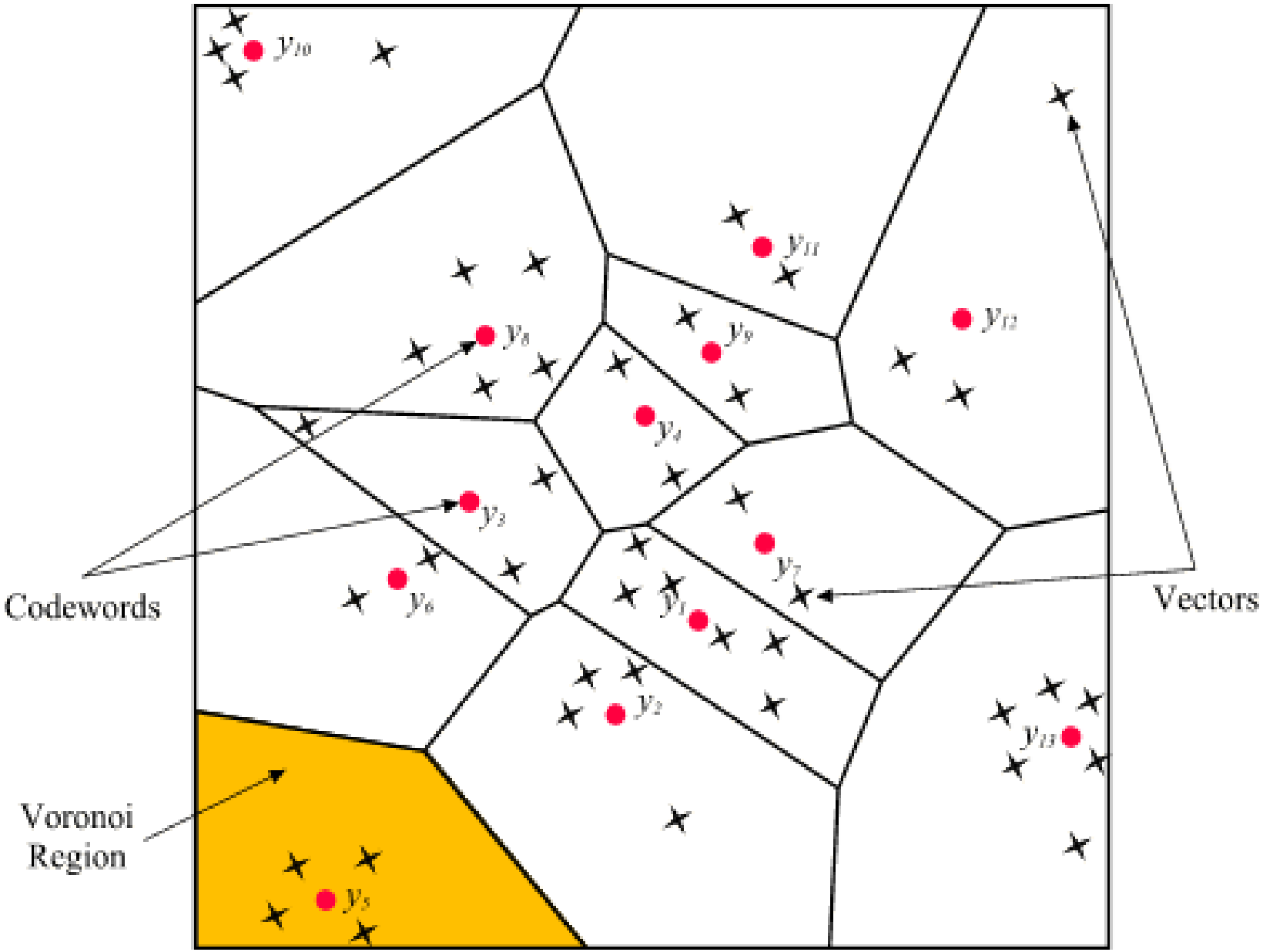
近似最近傍探索(ANN)と Vector Quantization (VQ)

多くのANNアルゴリズムは、ベクトル量子化（VQ）を使用しており、ベクトル空間を複数のグループに分割し、各グループを表す「コードワード」を定義し、それらのコードワードのみを検索します。

このVQテクニックはクエリー速度を劇的に向上させ、リレーショナル・データベースや全文検索エンジンにインデックス付けが不可欠であるように、多くのANNアルゴリズムに不可欠な部分になっています。

Vector Quantization については、次のページを参照ください。
https://www.mgasem.net/vectorquantization/vq.html

Vector Quantizationの例



ベクトル検索とキーワード検索とのハイブリッド検索

最初に、ベクトル検索のトレードオフを考えてみましょう。

ベクトルデータベースは、逆引きインデックスでのキーワード検索と比較して、見目は異なるが意味が同じものを見つける能力を提供しまが、メモリ使用量とレイテンシのコストは高くなります。また、確率論的であるため、速度を優先した場合絶対的なベストマッチを見逃す可能性があります。

例えば、特定のエラーコード「Error 504」や製品型番の検索を考えてみましょう。「Error 504」と「Error 503」はベクトル空間では意味的に近接しているかもしれないですが、現実には全く異なるものです。ベクトル検索はこうした検索では、失敗することが多くなります。

ハイブリッド検索
キーワード検索とベクトル検索の統合

ハイブリッド検索は、キーワードベースの検索とベクトルベースの検索という、2つの異なる検索手法をシームレスに組み合わせたアプローチです。

両方のアプローチの強みを活用することで、ハイブリッド検索はより正確で関連性の高い検索結果を提供します。

ただ、 2つの異なる検索手法をどのようにシームレスに組み合わせるのでしょう？　ハイブリッド検索の実装のポイントは、まさにその点にあります。

ハイブリッド検索の実装のポイント
二つの検索を共通の土台に立たせる

全く異なる手法であれば、二つの検索技術をシームレスに統合することは難しくなります。統合のポイントは、二つの検索を「共通の土台」に立たせるところにあります。

ハイブリッド検索の実装のポイントは、キーワード検索を特殊なベクトル検索としてとらえるところにあります。こうすることで、キーワード検索も、ベクトル検索という「共通の土台」に立たせることが可能になります。

sparse embedding

キーワードベース検索は、ベクトルの値の大部分がゼロで、ごく一部の値のみがゼロでないというembeddingに依存したベクトル検索と考えることができます。

このベクトルの値の大部分がゼロで、ごく一部の値のみがゼロでないというembeddingを、「スパース埋め込み（sparse embedding）」といいます。

このsparse embeddingは、後述のTF-IDF（Term Frequency-Inverse Document Frequency）アプローチに基づいたBM25などのアルゴリズムを使用して作成されます。

sparse embeddingを用いたキーワード検索

sparse embeddingを用いたキーワード検索の手順の概要をみておきましょう。コードレベルでの実装は、次回のセッションで紹介します。

- 前処理：
  検索クエリと文書を前処理し、関連するキーワードを抽出し、ノイズを除去します。

- スパース埋め込みの生成：
  コーパス内の各文書について、その文書内のキーワードの出現頻度を全文書における出現頻度と比較して、スパース埋め込みが生成されます。これは通常、BM25などのアルゴリズムを使用して行われます。

- クエリ埋め込み：
  検索クエリも、同じアルゴリズムを使用してスパース埋め込みに変換されます。

- 類似度計算：
  その後、コサイン類似度またはその他の類似度指標を用いて、クエリの埋め込みと各文書の埋め込みを比較し、各文書とクエリとの関連性を判定します。

- ランキング：
  文書は類似度スコアに基づいてランク付けされ、最も関連性の高い文書が検索結果の上位に表示されます。

次回のセッション

次回のセッションでは、Akash Chandrasekarの

“Hybrid Search: Combining Traditional Keyword-Based Search with Modern Vector Search”

https://medium.com/@csakash03/hybrid-search-is-a-method-to-optimize-rag-implementation-98d9d0911341 に基づいて、

sparse embeddingを用いたキーワード検索のプログラムを紹介しようと思います。


グローバルな検索の世界の変化
Sparse embedding
「知識のハブ」の変化を考える
embeddingの共有・蓄積・交換の未来

TF-IDF
Term Frequency-Inverse Document Frequency

以下の情報は、GoogleのAI検索に基づくものです。
とても便利だと思います。

TF-IDF（Term Frequency-Inverse Document Frequency）とは、文章中における単語の重要度を評価・数値化する統計的アルゴリズムです。主に自然言語処理における文書のベクトル化や、検索エンジン、レコメンデーション、テキストマイニングなどで広く活用されています。

TF-IDFを構成する2つの基本指標

TF-IDFは、以下の2つの指標を掛け合わせて算出します。

- TF（Term Frequency：単語頻度）
  - ある1つの文書内における「単語の出現頻度」です。
  - 意味：その文書内で多く使われている単語ほど重要度（TF値）が高くなります。

- IDF（Inverse Document Frequency：逆文書頻度）
  - 全文書集合における「単語の珍しさ（レア度）」です。
  - 意味：あらゆる文書に登場する一般的な単語（例：「これ」「する」など）の重要度を下げ、特定の文書にしか出ない珍しい単語の重要度（IDF値）を上げます。

TF-IDFの公式

$$\mathrm{TF} = \frac{\text{対象単語の出現回数}}{\text{その文書内の総単語数}}$$

$$\mathrm{IDF} = \log\left(\frac{\text{全文書数}}{\text{その単語を含む文書数}}\right)$$

$$\text{TF-IDF} = \mathrm{TF} \times \mathrm{IDF}$$

直感的な理解（具体例）

3つのニュース記事（文書A、B、C）を比較する場合を想定します。

- 「日本」という単語
  - 文書Aに多く出現（TFが高い）しても、文書Bや文書Cにも登場するため、全文書数3 / 含む文書数3となりIDFは0（あるいは極めて低く）になります。
  - 結果として、TF-IDF値は低くなり「この記事を特徴づけるキーワードではない」と判断されます。

- 「半導体」という単語
  - 文書Bにだけ何度も出現し、文書AとCには全く出ない場合、文書B内でのTFが高く、かつIDFも最高値になります。
  - 結果として、TF-IDF値は非常に高くなり、「文書Bは半導体に関する特徴的な記事である」と識別できます。

主な用途

- 情報検索（検索エンジン）：ユーザーが入力した検索クエリに対して、最も関連性の高い特徴を持つ文書を上位に表示します。
- 関連文書のレコメンド：記事や商品のテキスト情報をベクトル化し、コサイン類似度等を用いて「似ている内容の文書」を推薦します。
- 重要キーワード抽出：長文のテキストから、その文書を要約・代表する固有のキーワードを自動で抜き出します。
- SEOテキスト分析：競合コンテンツを分析し、自社のWebページに足りない重要語の過不足を可視化します。

知っておくべき限界（注意点）

- 文脈や語順の無視：単語の出現頻度だけを追うため、「犬が人を噛んだ」と「人が犬を噛んだ」を同じ意味の文章として扱ってしまいます。
- 同義語・類義語の未対応：「車」と「自動車」のように、意味が同じでも表記が異なる単語の関連性を自律的に捉えることができません。
- 文書の長さに影響される：極端に短い文書や長すぎる文書では、TF値のバランスが崩れ正しく評価できない場合があります。

ベクトル検索との関係

現在の最先端AI（大規模言語モデルやBERT)では文脈を捉える「埋め込みベクトル」が主流ですが、TF-IDFはその計算の軽快さと解釈性の高さから、現在も検索システムのベースライン（BM25アルゴリズム等）として強力に機能しています

BM25 (Best Matching 25)

以下の情報は、GoogleのAI検索に基づくものです。
とても便利だと思います。

BM25（Best Matching 25）は、検索エンジンが特定の検索クエリに対する文書の関連性を評価するために使用する、代表的なランキングアルゴリズムです。これはキーワードベースの検索における業界標準のベースラインであり、ElasticsearchやLuceneといったプラットフォームの中核を成しています。

中核となる計算式

BM25は、3つの主要な要素を用いて、クエリ用語の出現頻度に基づいて文書をランク付けします：

- 用語頻度（TF）： クエリ用語が文書内に何回出現するかをカウントします。
- 文書長正規化： 単に単語数が多いという理由だけで上位に表示されるのを防ぐため、長い文書に対してペナルティを課します。
- 逆文書頻度（IDF）： 一般的な単語（例：「the」、「and」）の重みを減らし、出現頻度が低く情報量の多い単語の重みを高めます。

BM25 vs. TF-IDF

BM25は、従来のTF-IDFが抱える根本的な課題である用語頻度の飽和に対処することで、その性能を向上させています。

- TF-IDF: 関連性スコアは直線的に増加します。あるキーワードが20回言及されている文書は、10回言及されている文書に比べて2倍の関連性スコアが与えられます。

- BM25: 関連性スコアは頭打ちになる。キーワードが5回出現するとスコアは大幅に上昇するが、20回と25回の出現回数による差は無視できるほど小さい。

調整可能なパラメータ

BM25では、開発者は2つの特定のパラメータを調整して検索動作を最適化することができます：

- （用語頻度飽和度）：用語頻度スコアがどのくらいの速さで頭打ちになるかを制御します。標準値は1.2から2.0の範囲です。値が低いほど、飽和が早く起こります。

- （文書長正規化）：文書長がスコアに与えるペナルティの度合いを制御します。標準値は 0.75です。これを 1.0に設定すると完全なペナルティが適用され、0に設定すると文書長は完全に無視されます。

現代的なユースケース

- 検索エンジン： Elasticsearch、OpenSearch、Apache Solrのデフォルトのテキストマッチング機能を支えています。
- ハイブリッド検索： 高密度ベクトル埋め込み（AI検索）と組み合わせることで、ニューラルネットワークでは見逃されがちなキーワードの完全一致を捕捉します。
- RAGパイプライン： RAG（Retrieval-Augmented Generation）システムにおいて、高速な第一段階の検索エンジンとして機能し、大規模言語モデル（LLM）に関連するコンテキストを取得します。

Sparse embedding プログラム サンプル

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Sample documents
documents = [
    "This is a sample document containing some keywords.",
    "Keywords are important for keyword-based search.",
    "Document analysis involves extracting keywords.",
    "Keyword-based search relies on sparse embeddings."
]

# Sample search query
query = "Keyword-based search"
```

```
# Step 1: Preprocessing (Skipped for brevity)
import re

def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()
    # Remove punctuation
    text = re.sub(r'[^¥w¥s]', '', text)
    return text

# Sample documents
documents = [
    "This is a sample document containing some keywords.",
    "Keywords are important for keyword-based search.",
    "Document analysis involves extracting keywords.",
    "Keyword-based search relies on sparse embeddings."
]

# Sample search query
query = "Keyword-based search"
```

```
# Preprocess documents
preprocessed_documents = [preprocess_text(doc) for doc in documents]

# Preprocess query
preprocessed_query = preprocess_text(query)

print("Preprocessed Documents:")
for doc in preprocessed_documents:
    print(doc)

print("¥nPreprocessed Query:")
print(preprocessed_query)
```

```
# Step 2: Sparse Embedding Generation (BM25)
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(documents)

# Step 3: Query Embedding
query_embedding = vectorizer.transform([query])

# Step 4: Similarity Calculation (Cosine Similarity)
similarities = cosine_similarity(X, query_embedding)

# Step 5: Ranking
ranked_indices = np.argsort(similarities, axis=0)[::-1].flatten()
ranked_documents = [documents[i] for i in ranked_indices]

# Output the ranked documents
for i, doc in enumerate(ranked_documents):
    print(f"Rank {i+1}: {doc}")
```

```
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

# Sample documents (represented as dense vectors)
document_embeddings = np.array([
    [0.634, 0.234, 0.867, 0.042, 0.249],
    [0.123, 0.456, 0.789, 0.321, 0.654],
    [0.987, 0.654, 0.321, 0.123, 0.456]
])

# Sample search query (represented as a dense vector)
query_embedding = np.array([[0.789, 0.321, 0.654, 0.987, 0.123]])

# Calculate cosine similarity between query and documents
similarities = cosine_similarity(document_embeddings, query_embedding)

# Rank documents based on similarity scores
ranked_indices = np.argsort(similarities, axis=0)[::-1].flatten()

# Output the ranked documents
for i, idx in enumerate(ranked_indices):
    print(f"Rank {i+1}: Document {idx+1}")
```







Part 3

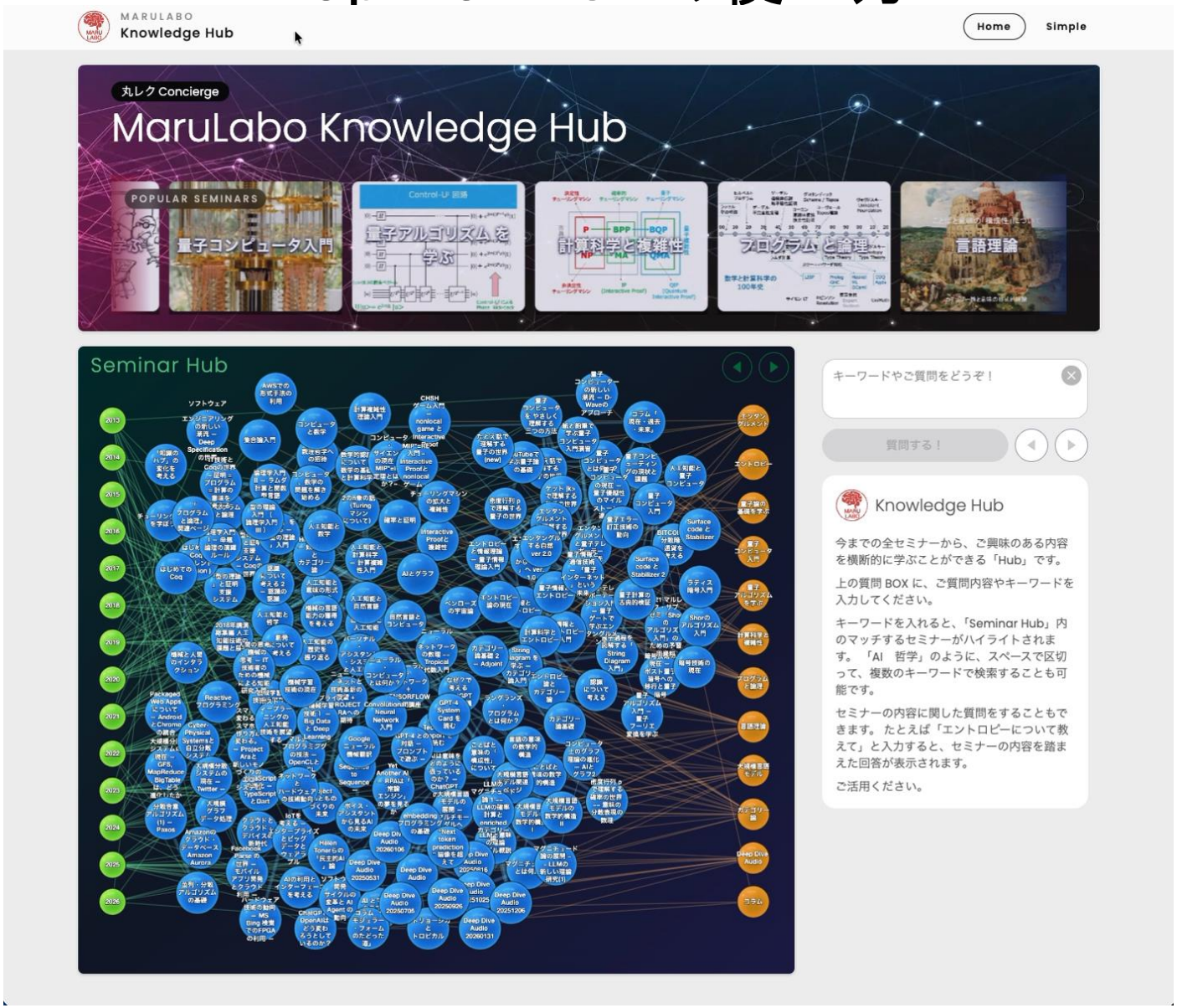
MaruLabo KnowledgeHub の紹介

MaruLabo KnowledgeHub (プロトタイプ版)とは何か

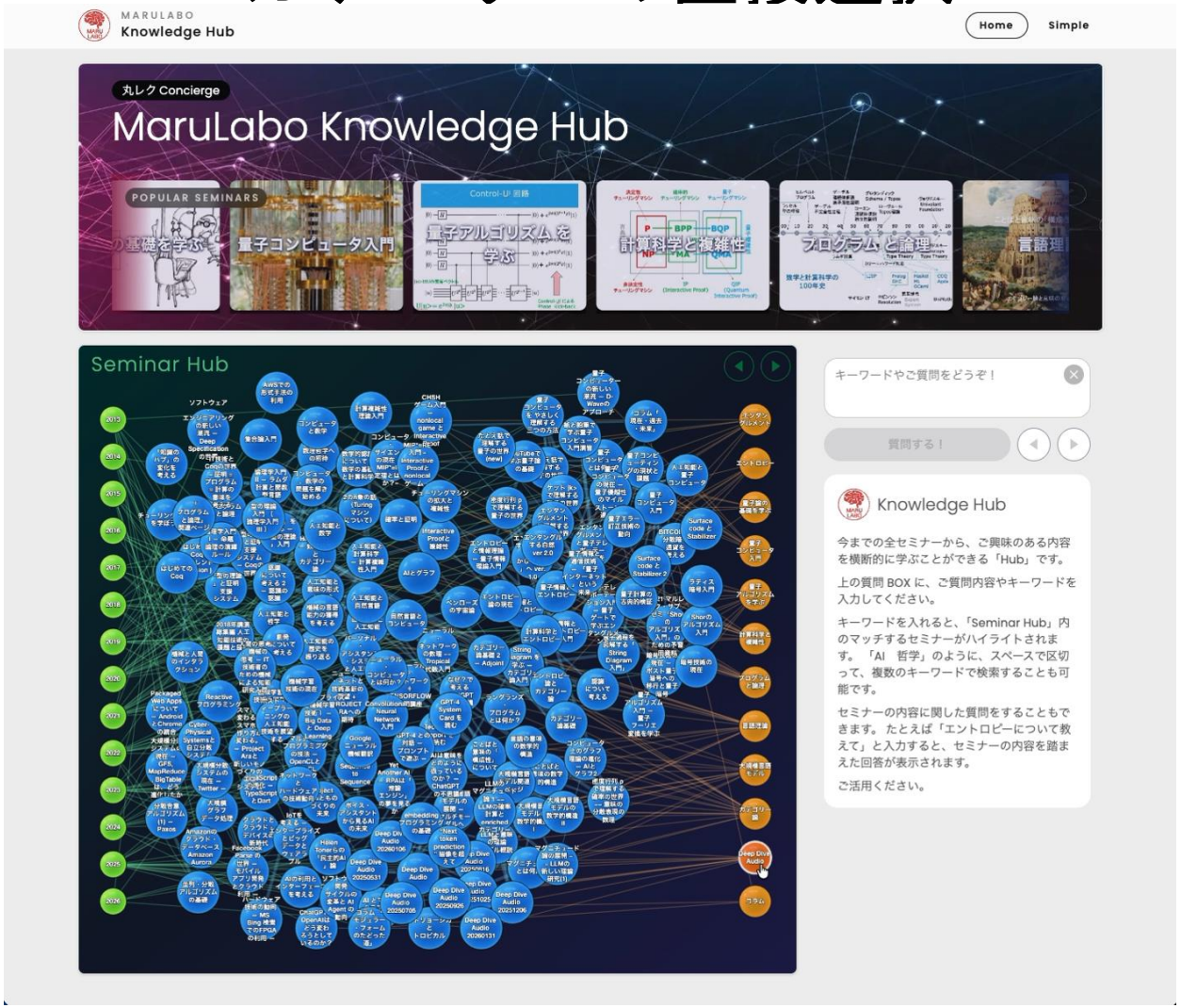
MaruLaboでは、マルレクのコンテンツを簡単にナビゲートできるシステム MaruLabo KnowledgeHub (M-KH が略称です)の開発を進めてきました。来週初めには、そのプロトタイプ版を公開する予定です。

今回のマルレクのテーマは、MaruLabo KnowledgeHub開発のコンセプトと結びついています。MaruLabo KnowledgeHubは、今回のセミナーのテーマの一つであるLLMのembedding技術を使った新しいタイプの検索技術の応用の実験でもあります。

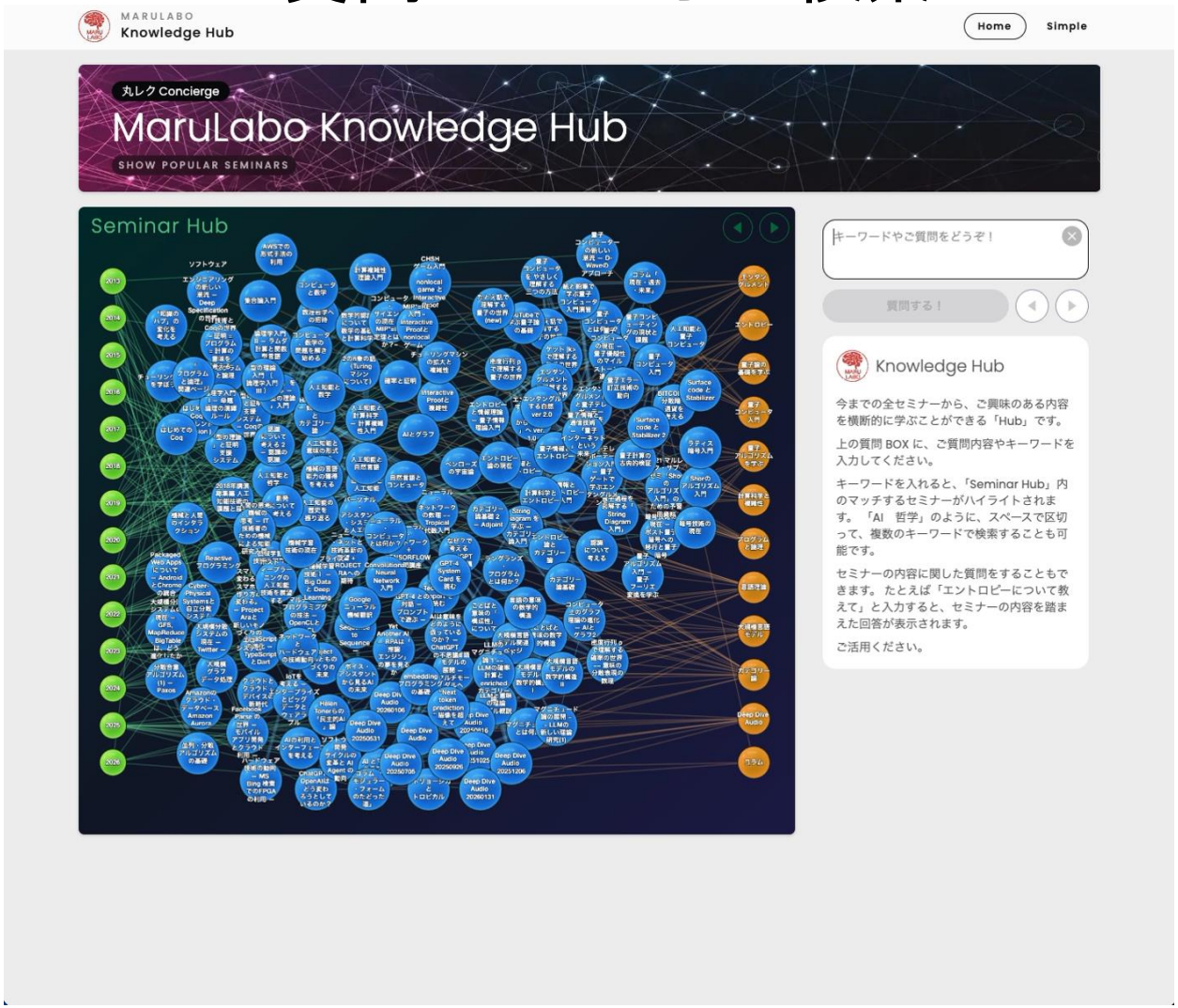
Top Banner の使い方



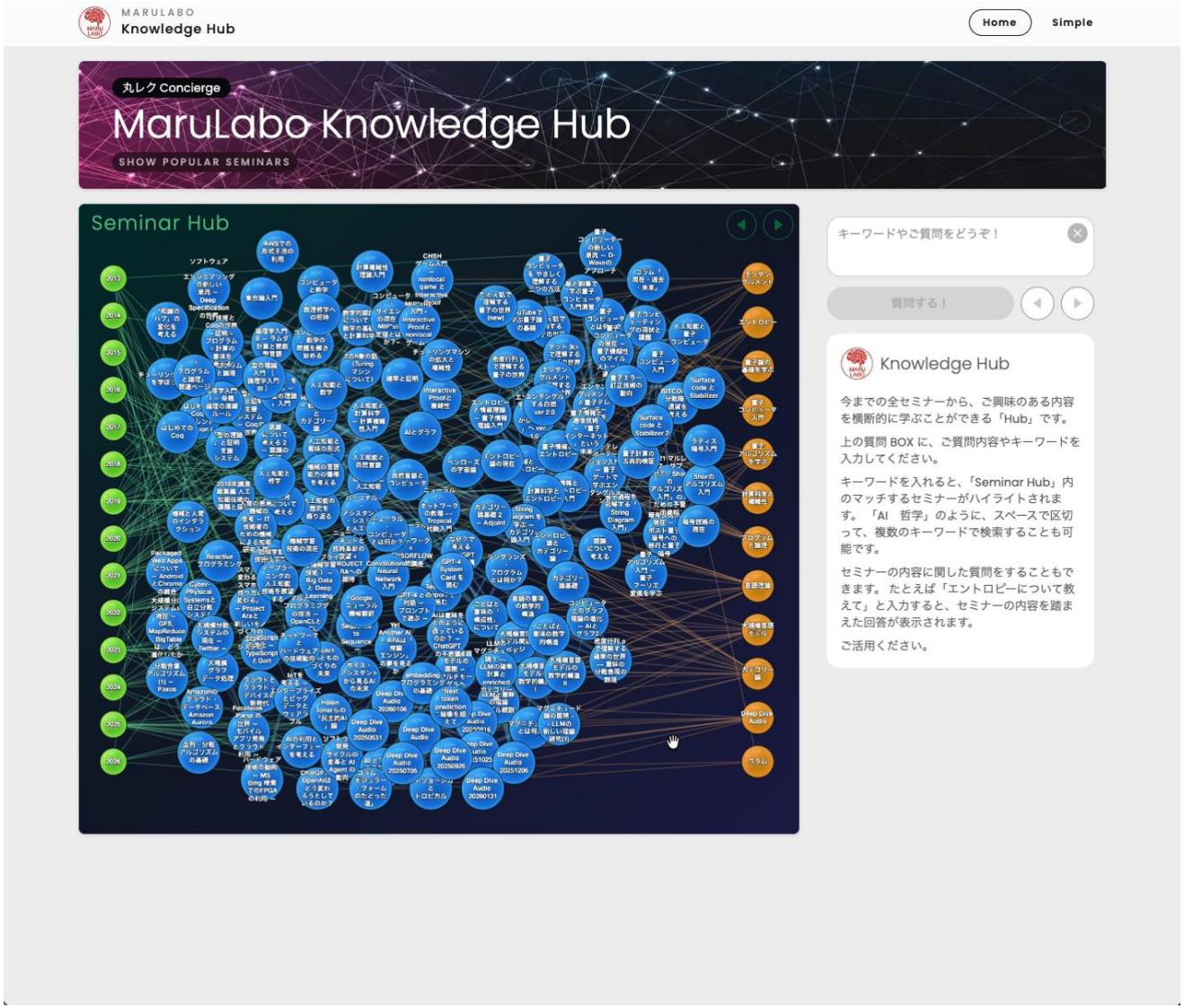
カテゴリーの直接選択


MARULABO
Knowledge Hub
Home
Simple
丸レク Concierge
MaruLabo Knowledge Hub
POPULAR SEMINARS
量子コンピュータ入門
量子アルゴリズムを学ぶ
計算科学と複雑性
プログラムと論理
言語理
Seminar Hub
キーワードやご質問をどうぞ！
質問する！
Knowledge Hub
今までの全セミナーから、ご興味のある内容を横断的に学ぶことができる「Hub」です。
上の質問 BOX に、ご質問内容やキーワードを入力してください。
キーワードを入れると、「Seminar Hub」内のマッチするセミナーがハイライトされます。「AI 哲学」のように、スペースで区切って、複数のキーワードで検索することも可能です。
セミナーの内容に関した質問をすることもできます。たとえば「エントロピーについて教えて」と入力すると、セミナーの内容を踏まえた回答が表示されます。
ご活用ください。

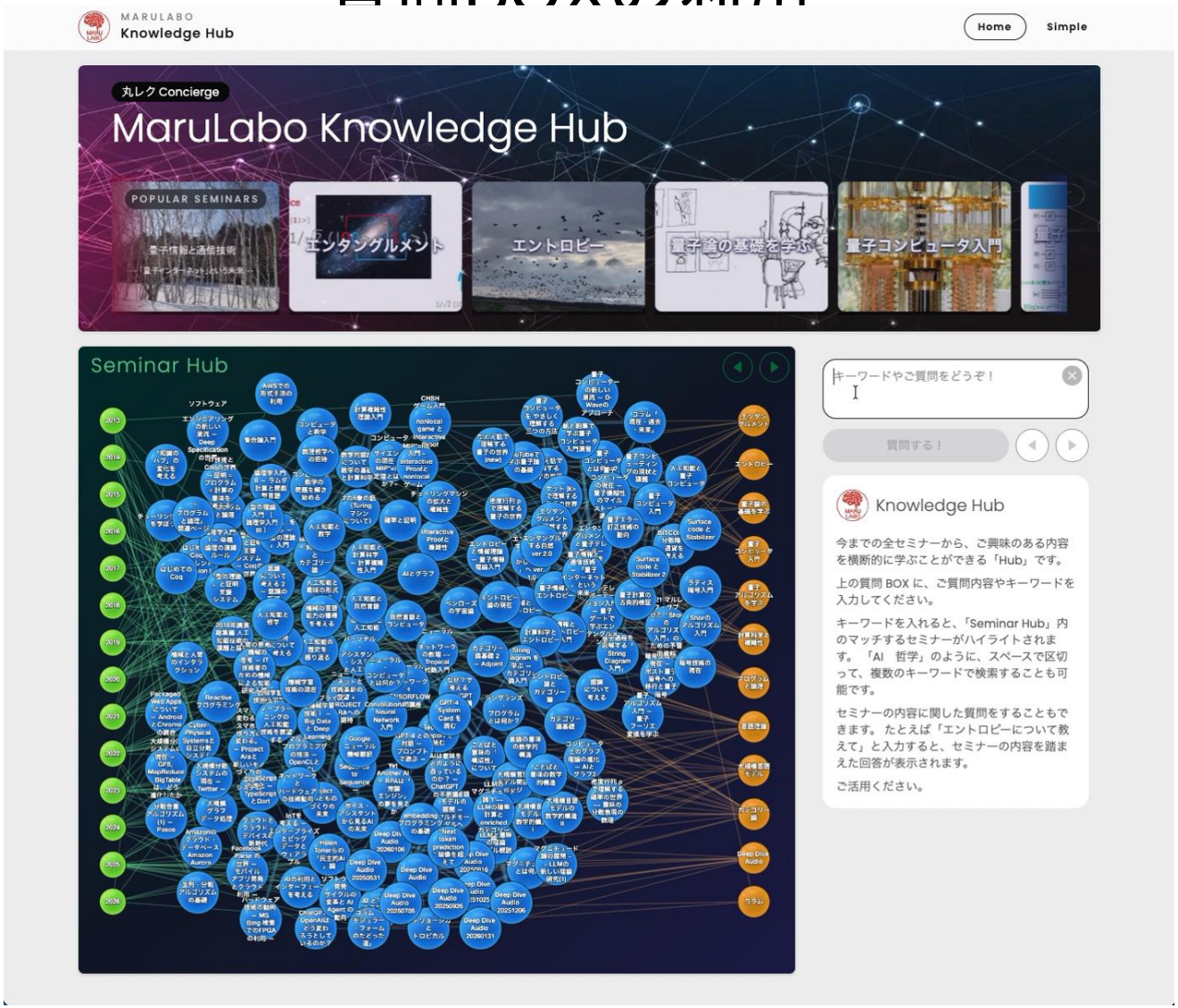
質問BOXからの検索


MARULABO
Knowledge Hub
Home
Simple
丸レク Concierge
MaruLabo Knowledge Hub
SHOW POPULAR SEMINARS
Seminar Hub
キーワードやご質問をどうぞ！
質問する！
Knowledge Hub
今までの全セミナーから、ご興味のある内容を横断的に学ぶことができる「Hub」です。
上の質問 BOX に、ご質問内容やキーワードを入力してください。
キーワードを入れると、「Seminar Hub」内のマッチするセミナーがハイライトされます。「AI 哲学」のように、スペースで区切って、複数のキーワードで検索することも可能です。
セミナーの内容に関した質問をすることもできます。 たとえば「エントロピーについて教えて」と入力すると、セミナーの内容を踏まえた回答が表示されます。
ご活用ください。

グラフの利用について


MARULABO
Knowledge Hub
Home
Simple
丸レク Concierge
MaruLabo Knowledge Hub
SHOW POPULAR SEMINARS
Seminar Hub
キーワードやご質問をどうぞ！
質問する！
Knowledge Hub
今までの全セミナーから、ご興味のある内容を横断的に学ぶことができる「Hub」です。
上の質問 BOX に、ご質問内容やキーワードを入力してください。
キーワードを入れると、「Seminar Hub」内のマッチするセミナーがハイライトされます。「AI 哲学」のように、スペースで区切って、複数のキーワードで検索することも可能です。
セミナーの内容に関した質問をすることもできます。 たとえば「エントロピーについて教えて」と入力すると、セミナーの内容を踏まえた回答が表示されます。
ご活用ください。

質問BOXの利用


MARULABO
Knowledge Hub
Home
Simple
丸レク Concierge
MaruLabo Knowledge Hub
POPULAR SEMINARS
量子情報と通信技術
エンタングルメント
エントロピー
量子論の基礎を学ぶ
量子コンピュータ入門
Seminar Hub
キーワードやご質問をどうぞ！
質問する！
Knowledge Hub
今までの全セミナーから、ご興味のある内容を横断的に学ぶことができる「Hub」です。
上の質問 BOX に、ご質問内容やキーワードを入力してください。
キーワードを入れると、「Seminar Hub」内のマッチするセミナーがハイライトされます。「AI 哲学」のように、スペースで区切って、複数のキーワードで検索することも可能です。
セミナーの内容に関した質問をすることもできます。たとえば「エントロピーについて教えて」と入力すると、セミナーの内容を踏まえた回答が表示されます。
ご活用ください。



