URL https://journal.ntt.co.jp/article/30171
DOI https://doi.org/10.60249/24111301

NTT人間情報研究所
特別研究員
増村 亮 Ryo Masumura

# 人間のように考え，人間のように知識を蓄積できるマルチモーダル基盤モデル「MediaGnosis（メディアグノシス）」

AI（人工知能）技術は進歩を続け，大規模言語モデルによる生成AIの普及によってより身近なものとなってきています．しかし今提供されているAIサービスは，顔認識機能，音声認識，翻訳・要約，文章生成等，特定の機能に特化しているものが主となっており，それを統合するようなAIサービスにはたどり着けていません．専門の知識を持った複数のAIをつなぎ合わせ，人の脳のように統合的に判断できる「MediaGnosis（メディアグノシス）」に取り組む増村亮特別研究員に，人間のような統合的なAIサービスの実現に向けたさまざまな課題や研究の心構えについてお話を伺いました．

◆PROFILE：2011年東北大学大学院 工学研究科 博士前期課程修了．同年，NTT 入社．2016年東北大学大学院 工学研究科 博士後期課程修了．音声認識や自然言語処理，動画像処理等のさまざまなメディア処理間で知識集約を行うことで，人間のように効率的に知識を蓄積してそれを活用する技術理論の研究開発に従事．日本音響学会 粟屋潔学術奨励賞，情報処理学会 山下記念研究賞，電子情報通信学会 情報・システムソサイエティ論文賞，言語処理学会 年次大会優秀賞などを受賞．博士（工学）．2019年より特別研究員．

## マルチモーダル基盤モデル「MediaGnosis（メディアグノシス）」がめざすもの

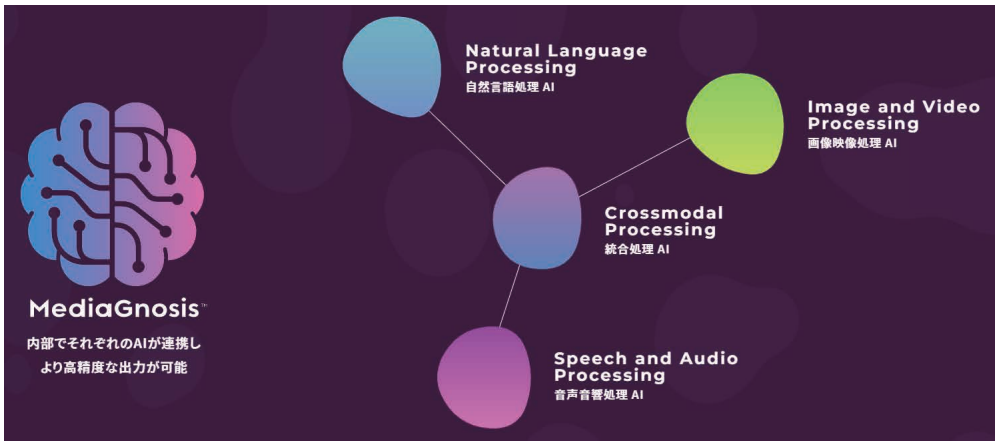
**■まず初めにご自身の研究テーマである「MediaGnosis」について教えてください．**

近年世の中では凄まじい勢いでAIブームが起きています．AIとはArtificial Intelligenceの略で「人工知能」を意味し，有名なものですと「ChatGPT」という大規模言語モデル（LLM）と呼ばれるAIがあります．ほかにも空港などで導入されている顔認識機能，人の話を理解する音声認識機能や，人の感情を読み取る機能などがあり，世間の注目を集めています．このAIは「人が実現するさまざまな知覚や知性を人工的に再現するもの」と位置付けられていますが，現状ではそこまでたどり着けておらず，特定の機能に特化して開発が進められています．例えば今のAIと人の脳を比較すると，人は音声を聴きながら映像をみて知識として取り入れて理解し，違う言語に翻訳することができます．「MediaGnosis（メディアグノシス）」は専門的な個々のAI機能に特化して進化させる研究ではなく，専門の知識を持った複数のAIをつなぎ合わせ，人の脳のように統合的に判断できるAIをつくる研究です（**図1**）．

AI分野では知識を蓄える“脳”を「モデル」と呼びますが，MediaGnosisはモデルを開発するのではなく「何でもできる」をめざした「マルチモーダル基盤モデル化」です．このマルチモーダル基盤モデル化は2つのマルチがポイントです．1つは「マルチモーダルデータの理解」で，個別に分離している機能に最適化された脳（モデル）を疎結合するのではなく，人のように1つの脳にさまざまな形式の知識を蓄え，より人らしく統合的に理解することを指します．もう1つは「マルチタスクをこなす」ことです．たくさんの知識を利用して，音声認識や感情理解など複数のAI機能を同時に駆動させながら人間らしい斬新な判断や，難しいタスクに取り組めることもめざします．このようにマルチモーダル・マルチタスクにまたがって多様な知識を利用し複数の機能を同時駆動させ，人間の脳に近いAIの研究開発がMediaGnosisです（**図2**）．

この研究を分かりやすくいえば漫画の世界で有名な，“何でもかなえてくれる猫型ロボット”や“10万馬力の人型ロボット”をイメージしてください．あのロボットたちはとても賢く，人とのコミュニケーションも卒なくこなして人を助けてくれます．私たちAI研究者のゴールは，あの世界で描かれているロボットの考え，対話する技術を開発することです．

今は音声認識，顔認識の機能やChatGPTもすごく良くなっています．しかし，それをただ組み合わせるだけであの有名な猫型ロボットができるのでしょうか．人が夢みるロボットたちをつくるためには，各機能が独立したただの機能ではなく人を理解して




- **MediaGnosisの由来**: あらゆるMedia（情報の記録）を，人間のように統合的にGnosis（知識）とし，それを基にDiagnosis（判断）する

図 1　MediaGnosis

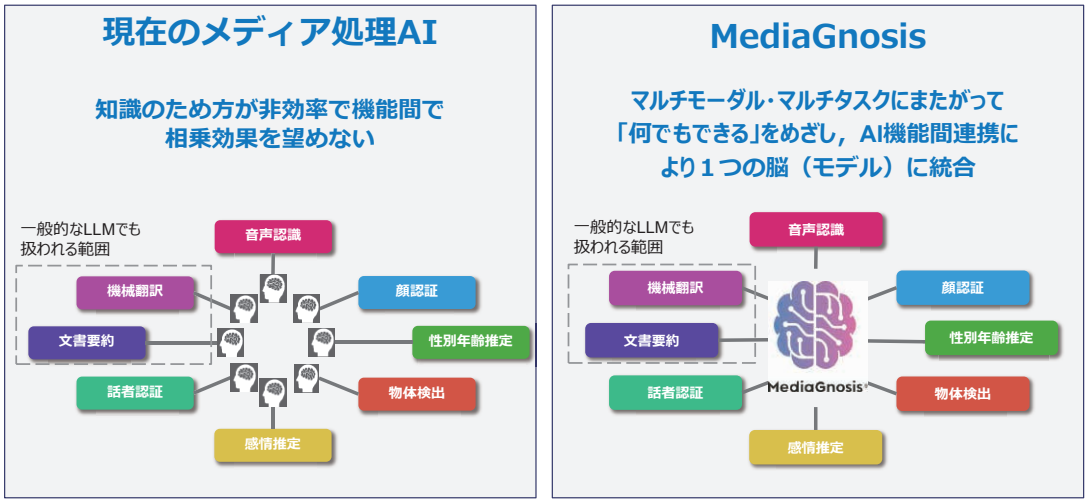



図 2　MediaGnosis の概要

人に寄り添う必要があります．各機能の精度が高まってきたからこそ，私たちはその先にある人間の脳のように情報を密につなげる研究開発をMediaGnosisと名付けて，日々研究に取り組んでいます．

### ■MediaGnosisならではの強みや特徴を教えてください．

AIは旬の研究テーマであり専門で研究されている方はとても多いのですが，その大半の方はメディアAIと呼ばれる“人の五感の１つ”に特化して研究を続けています．例えばLLMと呼ばれるAIは，ひと昔前までは機械翻訳や要約に関して研究していましたが，今ではChatGPTのように高い精度を持つAIに進化しました．私たちのMediaGnosisはLLMのように独自に進化するAI機能の延長線にあるのですが，五感の１つに特化せず五感全体を統合的に判断する研究であることが特徴です．ウィキペディアの隅々まで把握するような「深い言葉の理解」ではなく，言葉や映像などを「バランス良く統合して理解できる」サービスを強みとし，他社との違いを生み出しています．

MediaGnosisがめざす人の感情を理解し，他人への印象を理解できる高度な“人理解”の機能を開発することで，サービスを受ける側の満足度に影響をもたらします．そもそもコミュニケーションは与え手と受け手がいて初めて成り立ちますが，もし自分が誰かに謝らなければならないときに，謝る態度によって受け手が抱く印象や満足度は大きく変わります．私たちが開発したMediaGnosisのアプリでは，謝罪している映像や音声を基に相


図3　MediaGnosisによる分析結果

手に与える印象を統合的に分析できます．印象が悪くなる要因は人それぞれで，言い訳ととらえがちな言葉を多用したり，話すスピードやトーン，表情管理が適切ではなかったりします．それを各AIに蓄えられた情報を基に“脳”が分析し，項目ごとにフィードバックできます（**図3**）．

このようにMediaGnosisは多機能を同時に駆動し，人の深い部分を理解できるため，相手に不快感を与えないコミュニケーションの実現が可能になります．だからこそ，一般的に好まれるものだけでなく「この人はこういうタイプだから，こんなことを提供しよう」という高度な対応ができるようになり，サービスの幅が大きく広がります．

また，私たちのめざしている“すごく賢い脳”が完成したとしても，維持費が月額1億円もかかるようでは，サービスの実現は現実的ではありません．それを回避するためノートPCにも使われているCPUでも動かせるようなソフトウェアの軽量化も重視しており，独自で研究しています．MediaGnosisを導入する際，できる限り簡易かつ低コストで利用いただけることも強みの1つです．

NTT（Nippon Telegraph and Telephone Corporation）は社名が表すように電報・電話から始まる“コミュニケーションの無限の可能性”に挑む会社です．今はまだMediaGnosisのような統合分野の研究者は少ないのですが，私たちはNTTの研究チームとして，最先端のAI技術を用いて“人とのコミュニケーションを深めるサービスの開発”を先導していきたいと考えています．

**■MediaGnosis開発におけるご自身の研究内容について教えてください．**

MediaGnosisの研究開発は，私が立ち上げたテーマであり，所内のプロジェクトメンバーや関連会社とともに研究を進めています．このチームを立ち上げたとき，重要なのは専門性を持った各グループに適切な横串を刺すチーミングだと考えました．そのため，音声認識が得意なグループや言語理解が得意なグループなど，各専門分野に特化したグループを「ワンチーム」として座組みしたうえで，横串を刺して統合させることを重要視しています．もしそれぞれのグループが独自に進化したときにそのまま横串を刺すだけでは，プロジェクト全体としての進化は期待できません．私たちは，“統合した脳のモデルの仕組みを開発すること”を一貫してめざしており，最終目標の認識を統一しています．そのため音声認識を使って相手の話を100％理解させる個々の機能の成長ではなく，統合した環境の中で他の機能とデータを共有し，全体的にインテグレーションしてアップデートするという観点で，モデルをつくっています．

実際に，1つの精度の高い機能だけを稼動するよりも，同時に複数の機能を稼動させたほうが精度は高くなります．少ないリソース（容量）の中で同時に稼動させ，効率的かつ精度の高い判断をすることが望ましく，さらにどれくらいの比率で分配するかというバランスを考えることも重要になってきます（**図4**）．

AI開発の世界では，2つのデータモデルを結合させて生成するものを結合モデル，大量かつ多様なデータで訓練されたアプリケーションの基盤となるAIモデルのことを基盤モデルと呼びます．MediaGnosisはマルチモーダル基盤モデルというもので，マル

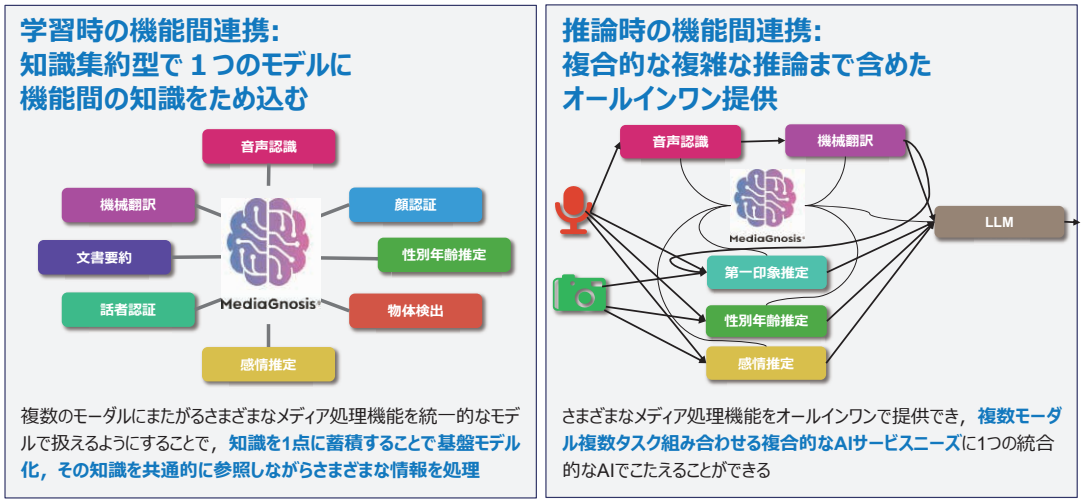



図４　MediaGnosisの機能間連携

チモーダルのモーダルは本や文章を読んだり言葉から学べること，映像を扱うなど“人の五感”を理解することを指します．人に近い脳をつくるためにはマルチタスクであることが必要で，音の言葉の成分や音の中の属性的な部分を理解し，複数のモーダルを同時に処理できることを指しています．例えば同じ意味の文章に対して返答するだけでも，怒っている男性が文章を早く出せ！と言っているのと，女性の方が優しく文章が欲しいです，と言う異なる情報を理解したうえで，適切な返答を判断できるようなマルチモーダル基盤モデルを開発しています．

### ■MediaGnosisの展望やゴールのビジョンなどを教えてください．

MediaGnosisの研究開発は私が特別研究員になったタイミングで立ち上げたもので，2019年から始動しています．そこから少しずつ統合する機能を増やして今のバージョンまで進化し，現在は人を理解すること（性別や年齢など答えが明確なもの）ができるまでになりました．しかし，私たちのめざす外的環境やコミュニティを理解するまでにはさらなる研究が必要です．

外的環境を理解する必要性をコールセンタで例えてみますと，オペレータは相手の話しだけではなく背景の環境音まで聞き取り，お客さまの状況を把握することが求められると聞きます．後ろで駅のアナウンスが流れている，複数人の声が聞こえる，テレビの音が聞こえるなど，その環境を理解して適切に応対することで円滑なコミュニケーションができるそうです．MediaGnosisも同じように環境を判断し，受け取り側に良い印象を与える表情や仕草，言葉の選び，話し方などが判断できることをめざしています．より多様な情報を扱い，一段深い情報を理解する機能を高めるためには，複数の情報を同時に取り込み，人のように判断することで，広範な分野で活用していただきたいと考えています（**図５**）．

しかし，この仕組みを世の中に理解してもらうためには，もう少し世の中がAIに対して成熟していく必要があります．ニュースなどで「AIに仕事を奪われるのでは」という言葉を耳にしますが，私たちAI技術開発者がめざすのはそこではありません．人の嫌がることをせず，人をサポートしてくれるAIの開発が目的です．MediaGnosisも人のように相手の感情や印象を理解して適切なコミュニケーションを取ることをめざしていますが，人に近づくことはできても完全な人の代替にはつながらないと考えています．

そもそもAIは人のようにあらなければならないのでしょうか．AIは人ではあらず，人のデータを蓄積するパートナー機能の１つととらえれば抵抗感も減少し，さまざまな場面において適応し活用できるはずです．そのうえで，サービスの面で人が求めるのはコミュニケーションであり，その感情を理解できるAIであるほうが，より世間に求められていると感じています．今はビジネスとして映像や文章だけの機能をサービスとして切り出して値付けをしている状態ですが，世間に認知してもらえなければ活躍の場は広がりません．そのためにはまずNTTにあるたくさんの事業会社にMediaGnosisを活用してもらい，精度や認知度を上げることが直近の目標です．すでにNTTコミュニケーションズ，NTTドコモ，NTT東日本・西日本でも導入や実証実験が行われていますが，ゆくゆくはNTT全体のエコシステムとしていきたいと考えており，その中央にはいつでもMediaGnosisがいることをめざしています．多くの場面で活用され，そこで得たデータを各社から受け取り，分析した結果をMediaGnosisの研究開発に還元してさらに成長させ，より良いサービスの提供をめざしています．

ほかにも関連性の深いものとしてNTTの掲げるIOWN（Innovative Optical and Wireless Network）構想の３つの柱の１つに“デジタルツインコンピューティング”があります．それは現実世界（リアル）のモノやヒトのツイン（双子）をデジタル世界上に構築することで，バーチャルの世界で代替するもので

■ 世の中のさまざまな場面で，MediaGnosisが当たり前に自然に埋め込まれている
**AI Anywhere の実現**をめざす
● **さまざまな場面で使われれば使われるほど成長**していくこともめざす

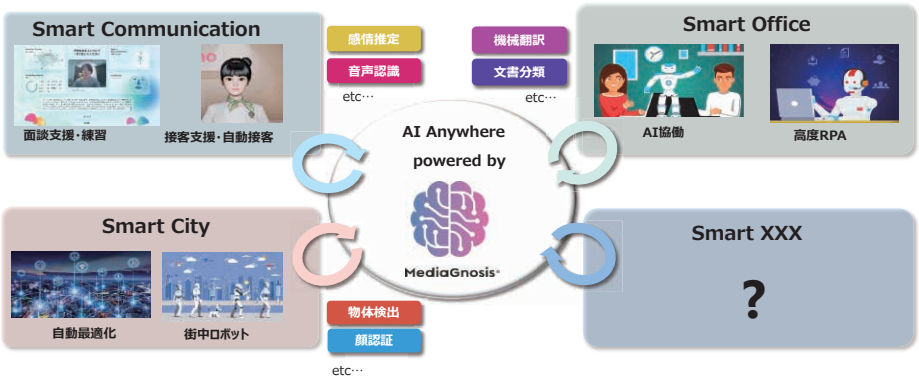



図 5　MediaGnosis のめざす世界

す．私の研究の骨子はデジタルツインコンピューティングにおける人の代替という部分に資する開発になります．MediaGnosisの「AI基盤としての強み」を最大限に引き上げるためには，高速かつ低消費電力なネットワークを基盤とするIOWN構想の実現が必要です．IOWN構想の革新的なネットワーク基盤が構築され，MediaGnosisがめざす，人のようなAIをつくることで，人の感情やコミュニケーションを理解する“人周辺の環境情報の理解”や，“組織や集団での役割や業務理解”などが進み，オフィス領域の協働作業はもちろん，介護や生活支援など幅広い人間の活動やさまざまな人に寄り添えるAIとして活躍してほしいと願っています．

## イノベーションを起こすための努力

**■研究における課題や，今後解決すべき問題などを教えてください．**

私の主な役割は横串を刺すことですが，具体的には複数の機能を1つのモデル（脳）として結合させる方法や，各機能のレベルが上がった際に他の機能レベルも追従して向上させる方法を研究しています．複数の機能をつなぐためには，脳として，分かれている他の機能を参照して活用する必要がありますが，その脳の統合がとにかく難しい課題です．例えば音声の感情推定部分のロジックを改変してデータ学習させたら，よく分からないけど翻訳の精度が下がってしまう，ということがあるのですが，このように1つの機能のバージョンアップによって全体のバランスが崩れるようなことは起こしたくありません．

それに加えアップデートの頻度も考えなければなりません．今は個々の機能に分かれてプール（蓄える）している複数のデータを中央にプールして一括学習させる仕組みのため，学習機能の更新や情報のバージョンアップは半年程度のスパンで行っています．人理解の部分では，最新データのアップデートはそこまで重要ではないため中央にプールする方法でも問題ないのですが，特定の人に寄り添う場合は話が変わってきます．例えば介護や看護ロボットなど生活の中に入るロボットに組み込まれるAIの場合，昨日のコミュニケーションを覚えたうえで対応できることが望ましいと考えています．現状はAIが昨日のニュースをすぐ取り込み知識として蓄えてアウトプットするのはとても難しいことです．しかし，小さなサイズでデータを蓄え，知識を構築できる脳の仕組みを構築できれば，半年ごとの学習スパンを短縮し，日々学習することも可能になります．高く掲げたゴールをめざして進んでいますが，現状は何かを学習したらどこかに性能劣化が起きることがあります．そこを乗り越えた先にめざす未来があると考えています．

**図 6** の画像の3つの「エンジン」をリンキングするのが横串です．重要なのは異なる事象や情報を多面的に理解し，認識対象が同一であるとリンクさせていくことで，それをいかに細かくリンキングできるかということにも取り組んでいます．

また，UI（User Interface）についてもよく質問をいただきます．もしブラウザに限定してしまうとPCの前でしかサービスの提供はできません．現時点ではMediaGnosisも利用環境や導入のしやすさからブラウザを選択することが多いですが，ブラウザに限らずロボットやタッチディスプレイ，監視（管理）カメラなど，比較的どれにでも対応できるようにAPI（Application Programming Interface）ベースで開発を進めています．もちろん今後どのインタフェースが主流になるのか見据える必要はありますが，インタフェースによって精度が変わってしまうことがないように心掛けています．

ほかにも，人種によって不適切な動きや言動は変わるといわれており，人を嫌な気持ちにさせてしまうAIであってはいけませんので，このような個別対応もクリアすべき問題です．そもそもAI

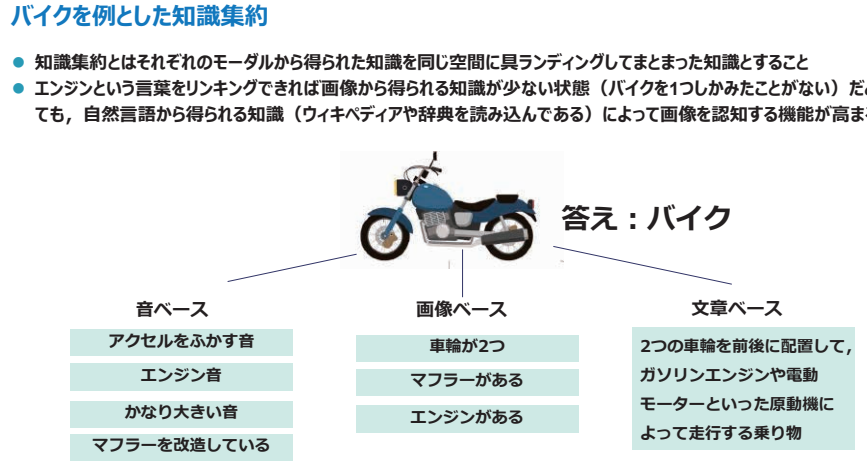



図6　知識集約とは

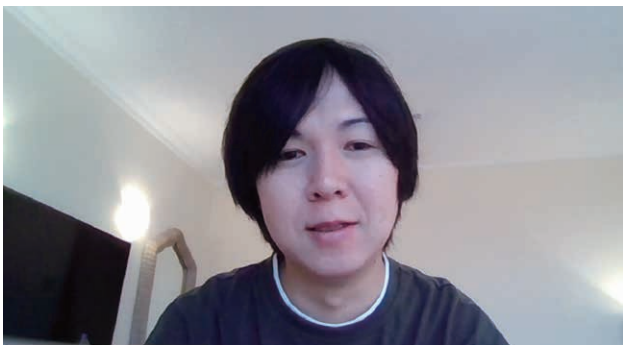
というものは“リリースが完全版”ではなく，“人によって成長させてもらう”ものです．使ってみたらこんな問題が出てきたとフィードバックを受け，それを改善して精度を上げるものなのです．どのようなサービスでも最初は受け入れてはもらえないものです．ネットショッピングも，今は当たり前になりましたが，最初は抵抗感を示す人が多く，少しずつ利用者が増えることで抵抗感が減り普及していきました．少しずつ世間がAIの特性を理解し，利用してもらえる社会になったときに，より精度の高いMediaGnosisでのサービス提供ができるように研究開発を進めていきます．

**■最後に，研究者・学生・ビジネスパートナーの方々へメッセージをお願いします．**

世の中にイノベーション思考という言葉がありますが，研究開発のためには本質の課題やテーマを深く考えて創造することが必要です．研究者の中には「次の国際会議に通せる研究」という目的意識で研究を進めてしまう場合がありますが，それでは本質の課題に深く向き合えないのではないかと思っています．やはり研究者としてはイノベーションを生み出したいもので，それは本質的な課題を創造的な発明で解決し，この世界をより幸せな未来にしていくことだと考えています．その期待にこたえるかたちで，学生や若い研究者はイノベーション思考であってほしいと願います．

また，大きな未来を掲げたときに，その未来に至るアプローチを逆算しなければなりません．研究者の中には天才型と努力型がいますが，天才型の人は好きなことをすればよいと思います．しかし努力型の人は逆算思考を持ち，この未来に到達するための具体的な目標を設定して進めることが大切です．研究者としてイノベーション思考で目標を設定し，その目標のために逆算思考で考えるということを意識してみるとよいと思います．

そして，ビジネスパートナーの方にとってMediaGnosisがより良いサービスとなるためには，まずは使っていただくことが第一になります．私たちの技術は広い分野で活用でき，そこを強みにすべく研究を進めています．MediaGnosisが利便性と汎用性のあるサービスになるためには，利用した結果のフィードバックを受け改善するサイクルが重要です．多くのビジネスパートナーの方々と検証サイクルを繰り返して精度を上げていき，利用者の期待にこたえていきたいと考えています．そのためビジネスパートナーの方々にはぜひ広い分野で使い続けていただき，MediaGnosisの期待や未来像を私たちとともに成長させていく，そのような相乗効果を生み出す関係が築ければと考えています．



（今回はリモートにてインタビューを実施しました）