(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2019-526851
(P2019-526851A)
(43)公表日　令和1年9月19日(2019.9.19)

(51)Int.Cl.		F I		テーマコード（参考）
G06N 20/00	(2019.01)	G06N	20/00	5L099
G16H 10/00	(2018.01)	G16H	10/00	

審査請求 有　予備審査請求 未請求　（全 51 頁）

(21)出願番号　特願2019-502045 (P2019-502045)
(86)(22)出願日　平成29年7月17日 (2017.7.17)
(85)翻訳文提出日　平成31年2月20日 (2019.2.20)
(86)国際出願番号　PCT/US2017/042356
(87)国際公開番号　WO2018/017467
(87)国際公開日　平成30年1月25日 (2018.1.25)
(31)優先権主張番号　62/363,697
(32)優先日　平成28年7月18日 (2016.7.18)
(33)優先権主張国・地域又は機関
米国 (US)

(71)出願人　513276204
ナント　ホールディングス　アイピー　エルエルシー
Ｎａｎｔ　Ｈｏｌｄｉｎｇｓ　ＩＰ，　ＬＬＣ
アメリカ合衆国　カリフォルニア州　９０２３２　カルバーシティ　ジェファーソン　ブールバード　９９２０
９９２０　Ｊｅｆｆｅｒｓｏｎ　Ｂｌｖｄ　Ｃｕｌｖｅｒ　Ｃｉｔｙ　Ｃａｌｉｆｏｒｎｉａ　９０２３２　Ｕ．Ｓ．Ａ．

最終頁に続く

(54)【発明の名称】分散型機械学習システム、装置、および方法

(57)【要約】
分散型オンライン機械学習システムを提供する。想定されるシステムは、それぞれがローカルプライベートデータを有する多くのプライベートデータサーバを含む。研究者は、プライベートデータの匿名化を要求することなくまたはプライベートデータを無許可のコンピューティングシステムに晒すことなく、関連するプライベートデータサーバが、機械学習アルゴリズムの実装をそれらのローカルプライベートデータでトレーニングすることを要求できる。また、プライベートデータサーバは、実際のデータのデータ分布に従って合成データまたはプロキシデータを生成する。サーバは、プロキシデータを使用してプロキシモデルをトレーニングする。プロキシモデルがトレーニング済み実モデルと十分に類似している場合、プロキシデータ、プロキシモデルパラメータ、または他の学習された知識を１つまたは複数の非プライベート演算デバイスに送信できる。多くのプライベートデータサーバから学習された知識は、プライベートデータを公開することなく、１つ以上のトレーニング済みグローバルモデルに集約できる。

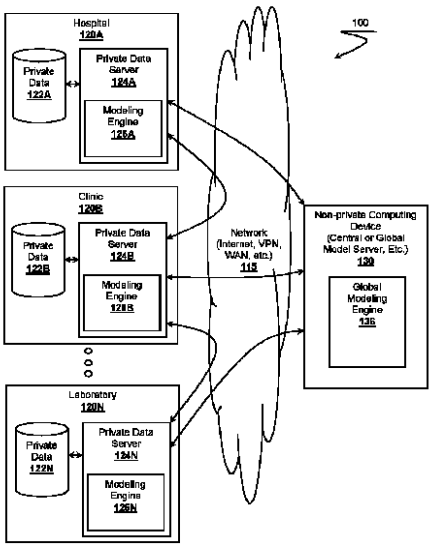


Figure 1

【特許請求の範囲】
【請求項１】
分散型機械学習システムであって、
それぞれがローカルプライベートデータへのアクセスを有しおよび少なくとも１つのモデリングエンジンを有する複数のプライベートデータサーバを備え、前記複数のプライベートデータサーバは、ネットワークを介して、少なくとも１つの非プライベート演算デバイスに通信可能に接続されており、
各前記プライベートデータサーバは、非一時的コンピュータ可読メモリに格納された少なくとも１つのプロセッサソフトウェア命令による実行に応じて、その少なくとも１つのモデリングエンジンに、
前記ローカルプライベートデータの少なくとも一部から、機械学習アルゴリズムの実装に従って、トレーニング済み実モデルを作成するためのモデル指示を受信するステップ、
前記機械学習アルゴリズムの実装を前記ローカルプライベートデータでトレーニングすることによって、前記モデル指示に従っておよび前記ローカルプライベートデータの少なくとも一部の関数として、トレーニング済み実モデルパラメータを含む前記トレーニング済み実モデルを作成するステップ、
前記ローカルプライベートデータから、前記トレーニング済み実モデルを作成するために使用される前記ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成するステップ、
前記複数のプライベートデータ分布に従ってプロキシデータのセットを生成するステップ、
前記機械学習モデルのタイプを前記プロキシデータのセットでトレーニングすることによって、前記プロキシデータのセットから、プロキシモデルパラメータを含むトレーニング済みプロキシモデルを作成するステップ、
前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数としてモデル類似性スコアを算出するステップ、
前記ネットワークを介して、前記モデル類似性スコアの関数として、前記プロキシデータのセットを少なくとも１つの非プライベート演算デバイスに送信するステップ、
を実行させるシステム。
【請求項２】
請求項１記載のシステムであって、前記ローカルプライベートデータは、ローカルプライベート健康管理データを含むシステム。
【請求項３】
請求項２において、前記ローカルプライベート健康管理データは、患者固有のデータを含むシステム。
【請求項４】
請求項１において、前記ローカルプライベートデータは、ゲノムデータ、全ゲノム配列データ、全エキソソーム配列データ、プロテオームデータ、プロテオミクス経路データ、ｋ－ｍｅｒデータ、ネオエピトープデータ、ＲＮＡデータ、アレルギー情報、遭遇データ、治療データ、転帰データ、予約データ、注文データ、請求コードデータ、診断コードデータ、結果データ、治療反応データ、腫瘍反応データ、人口統計データ、投薬データ、バイタルサインデータ、支払者データ、薬物研究データ、薬物反応データ、経時的研究データ、バイオメトリックデータ、財務データ、所有権データ、電子カルテデータ、研究データ、人材データ、パフォーマンスデータ、分析結果データ、または事象データを含むデータの少なくとも１つを含むシステム。
【請求項５】
請求項１において、前記ネットワークは、無線ネットワーク、パケット交換ネットワーク、インターネット、イントラネット、仮想プライベートネットワーク、セルラネットワーク、アドホックネットワーク、およびピアツーピアネットワークの少なくとも１つを含むシステム。

【請求項６】
請求項１において、前記少なくとも１つの非プライベート演算デバイスは、前記トレーニング済み実モデルが作成されたローカルプライベートデータに対する権限がない前記複数のプライベートデータサーバのうちの異なる１つであるシステム。
【請求項７】
請求項１において、前記少なくとも１つの非プライベート演算デバイスは、グローバルモデルサーバを含むシステム。
【請求項８】
請求項８において、前記グローバルモデルサーバは、前記複数のプライベートデータサーバのうちの少なくとも２つからのプロキシデータのセットを集約するように構成され、グローバルモデルを前記プロキシデータのセットでトレーニングするように構成されているシステム。
【請求項９】
請求項１において、各前記プライベートデータサーバは、前記ローカルプライベートデータを格納するローカルストレージシステムに通信可能に接続されているシステム。
【請求項１０】
請求項９において、前記ローカルストレージシステムは、ＲＡＩＤシステム、ファイルサーバ、ネットワークアクセス可能なストレージデバイス、ストレージエリアネットワークデバイス、ローカルコンピュータ可読メモリ、ハードディスクドライブ、光ストレージデバイス、テープドライブ、テープライブラリ、およびソリッドステートディスクの少なくとも１つを含むシステム。
【請求項１１】
請求項９において、前記ローカルストレージシステムは、ローカルデータベース、ＢＡＭサーバ、ＳＡＭサーバ、ＧＡＲサーバ、ＢＡＭＢＡＭサーバ、および臨床オペレーティングシステムサーバの少なくとも１つを含むシステム。
【請求項１２】
請求項１において、前記モデル指示は、ローカルコマンド、リモートコマンド、実行可能ファイル、プロトコルコマンド、および選択されたコマンドの少なくとも１つを含むシステム。
【請求項１３】
請求項１において、前記複数のプライベートデータ分布の分布は、ガウス分布、ポアソン分布、ベルヌーイ分布、ラデマッハ分布、離散分布、二項分布、ゼータ分布、ガンマ分布、ベータ分布、およびヒストグラム分布の少なくとも１つに従うシステム。
【請求項１４】
請求項１において、前記複数のプライベートデータ分布は、前記トレーニング済み実モデルパラメータと前記ローカルプライベートデータから導出された固有値に基づくシステム。
【請求項１５】
請求項１において、前記プロキシデータのセットは、前記トレーニング済み実モデルパラメータと前記ローカルプライベートデータから導出された固有ベクトルの組み合わせを含むシステム。
【請求項１６】
請求項１５において、前記プロキシデータは、前記固有ベクトルの線形結合を含むシステム。
【請求項１７】
請求項１５において、前記固有ベクトルは、固有の患者、固有のプロファイル、固有の薬剤、固有の健康記録、固有のゲノム、固有のプロテオーム、固有のＲＮＡプロファイル、および固有の経路の少なくとも１つを含むシステム。
【請求項１８】
請求項１において、前記トレーニング済み実モデルは、分類アルゴリズム、ニューラル

ネットワークアルゴリズム、回帰アルゴリズム、決定木アルゴリズム、クラスタリングアルゴリズム、遺伝的アルゴリズム、教師あり学習アルゴリズム、半教師あり学習アルゴリズム、教師なし学習アルゴリズム、および深層学習アルゴリズムを含む機械学習アルゴリズムの少なくとも１つの実装に基づくシステム。

【請求項１９】

請求項１において、前記トレーニング済み実モデルは、サポートベクターマシン、最近傍アルゴリズム、ランダムフォレスト、リッジ回帰、Ｌａｓｓｏアルゴリズム、ｋ－ｍｅａｎｓクラスタリングアルゴリズム、スペクトルクラスタリングアルゴリズム、平均シフトクラスタリングアルゴリズム、非負行列因数分解アルゴリズム、エラスティックネットアルゴリズム、ベイズ分類アルゴリズム、ＲＡＮＳＡＣアルゴリズム、および直交マッチング追跡アルゴリズムを含む機械学習アルゴリズムの少なくとも１つの実装に基づくシステム。

【請求項２０】

請求項１において、前記モデル指示は、前記プライベートデータサーバの外部で作成されたベースラインモデルから前記トレーニング済み実モデルを作成するための指示を含むシステム。

【請求項２１】

請求項２０において、前記ベースラインモデルは、グローバルトレーニング済み実モデルを含むシステム。

【請求項２２】

請求項２１において、前記グローバルトレーニング済み実モデルは、少なくとも部分的に、前記複数のプライベートデータサーバのうちの少なくとも２つからのプロキシデータのセットでトレーニングされるシステム。

【請求項２３】

請求項１において、前記類似性スコアは、前記プロキシモデルの交差検証に基づいて判定されるシステム。

【請求項２４】

請求項２３において、前記交差検証は、前記プロキシデータの一部に対する内部交差検証を含むシステム。

【請求項２５】

請求項２３において、前記交差検証は、前記ローカルプライベートデータの内部交差検証を含むシステム。

【請求項２６】

請求項２３記載において、前記交差検証は、前記複数のプライベートデータサーバのうちの異なる１つによるそのローカルプライベートデータでの外部クロス検証を含むシステム。

【請求項２７】

請求項１において、前記類似性スコアは、前記プロキシモデルの正確度と前記トレーニング済み実モデルの正確度との差を含むシステム。

【請求項２８】

請求項１において、前記類似性スコアは、前記トレーニング済み実モデルパラメータと前記プロキシモデルパラメータから算出されたメトリック距離を含むシステム。

【請求項２９】

請求項１において、前記プロキシデータは、前記モデル類似性スコアの関数が少なくとも１つの送信基準を満たすときに送信されるシステム。

【請求項３０】

請求項２９において、前記少なくとも１つの送信基準は、類似性スコアに関して、閾値条件、多値条件、値の変化条件、傾向条件、人的命令条件、外部要求条件、および時間条件の少なくとも１つを含むシステム。

【請求項３１】

請求項１において、前記モデリングエンジンは、前記トレーニング済み実モデルを新しいローカルプライベートデータで更新するように更に構成されるシステム。
【請求項３２】
コンピュータデバイスにより実装される分散機械学習の方法であって、
プライベートデータサーバによって、前記プライベートデータサーバにローカルなローカルプライベートデータの少なくとも一部から、機械学習アルゴリズムの実装に従って、トレーニング済み実モデルを作成するためのモデル指示を受信するステップ、
機械学習エンジンによって、前記機械学習アルゴリズムの実装を前記ローカルプライベートデータでトレーニングすることによって、前記モデル指示に従って、および前記ローカルプライベートデータの少なくとも一部の関数として、トレーニング済み実モデルパラメータを含む前記トレーニング済み実モデルを作成するステップ、
前記機械学習エンジンによって、前記ローカルプライベートデータから、前記トレーニング済み実モデルを作成するために使用される前記ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成するステップ、
前記機械学習エンジンによって、前記プライベートデータ分布から、前記複数のプロキシデータ分布の複製を可能にする顕著なプライベートデータ特徴を識別するステップ、
前記機械学習エンジンによって、ネットワークを介して、前記顕著なプライベートデータ特徴を非プライベート演算デバイスに送信するステップ、
を含む方法。
【請求項３３】
請求項３２において、前記顕著なプライベートデータ特徴は、プロキシデータのセットを含む方法。
【請求項３４】
請求項３２において、前記複数のプライベートデータ分布および顕著なプライベートデータ特徴の少なくとも１つに従ってプロキシデータのセットを生成するステップを更に含む方法。
【請求項３５】
請求項３４において、前記機械学習アルゴリズムの実装のタイプを前記プロキシデータのセットでトレーニングすることによって、前記プロキシデータのセットから、プロキシモデルパラメータを含むトレーニング済みプロキシモデルを作成するステップを含む方法。
【請求項３６】
請求項３５において、前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数として、前記トレーニング済みプロキシモデルのモデル類似性スコアを算出するステップを更に含む方法。
【請求項３７】
請求項３６において、前記モデル類似性スコアに基づいて、前記プロキシデータのセットを集約グローバルモデルに集約するステップを更に含む方法。
【請求項３８】
ローカルプライベートデータにアクセスするように構成され、少なくとも１つのモデリングエンジンを含むプライベートデータサーバを使用してプロキシデータを生成するコンピュータ実装方法であって、前記少なくとも１つのモデリングエンジンは、
前記プライベートデータから、機械学習アルゴリズムを使用してトレーニング済み実モデルを作成し、
前記ローカルプライベートデータの少なくとも一部から、ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成し、
前記複数のプライベートデータ分布に基づいて、プロキシデータのセットを生成し、
前記プロキシデータのセットから、前記機械学習アルゴリズムを使用して、トレーニング済みプロキシモデルを作成する、
ように構成されている方法。

【発明の詳細な説明】
【技術分野】
【０００１】
＜関連出願への相互参照＞
本出願は、米国特許法第１１９条に基づき、Ｓｚｅｔｏにより２０１６年7月１８日に出願された、米国仮特許出願第６２／３６３，６９７号、発明の名称「Distributed Machine Learning Systems, Apparatus, and Methods」の優先権を主張し、この内容全体は、参照により本明細書に援用される。
【０００２】
本発明は、分散型機械学習技術に関する。
【背景技術】
【０００３】
背景技術の説明は、本発明の主題を理解するために有用となる情報を含む。なお、この説明は、本明細書に提供する全ての情報が先行技術または現在特許請求している発明の主題に関連することを認めるものではなく、あるいは、具体的または暗黙的に言及する刊行物が先行技術であることを認めるものではない。
【０００４】
アクセス性が高く費用対効果が高い機械学習プラットフォーム（テンソルフロー（TensorFlow）を含むグーグル（Google）社の人工知能、アマゾン（Amazon）社の機械学習、マイクロソフト（Microsoft）社のアジュール（Ａｚｕｒｅ）機械学習、オープンＡＩ（OpenAI）、ＳｃｉＫｉｔ－Ｌｅａｒｎ、Ｍａｔｌａｂ等）の近年の成長によって、データアナリストは、大規模データセットの自動分析を実行するための多数の既製の（off-the-shelf）オプションを有するようになった。更に、機械学習プラットフォームの成長と並行して、ターゲットデータセットのサイズも大きくなった。例えば、ヤフー（Yahoo!）は、テラバイトの桁のサイズを有する幾つかの大規模なデータセットを公開しており、癌ゲノムアトラス（Cancer Genome Atlas：ＴＣＧＡ）データポータルは、大量の臨床情報およびゲノム特性データへのアクセスを提供している。これらの事前作成されたデータセットは、データアナリストが容易に利用できる。
【０００５】
しかしながら、研究者は、進行中の研究のためにデータセットを編集する際、特に実地データ（in-the-field data）を使用して興味深い予測を生成できるトレーニング済み機械学習モデルを構築しようとする際に、障害に遭遇することが多い。大きな障害の１つは、研究者が必要とするデータにアクセスできないことが多いということである。例えば、研究者が、複数の病院の電子医療記録データベースに格納されている患者データから、トレーニング済みモデルを構築するシナリオについて検討する。研究者は、プライバシの制限またはＨＩＰＡＡのコンプライアンスのために、各病院の患者データにアクセスする権限を有していない可能性がある。所望のデータセットを編集するために、研究者は、病院にデータを要求しなければならない。病院がその要求に従う場合、病院は、データを研究者に提供する前に、特定の患者への紐づけを削除するために、そのデータを匿名化しなければならない。しかしながら、匿名化により、機械学習アルゴリズムのトレーニングに役立ち、したがって、データ内の新しい関係を発見する機会を提供したり、価値予測特性を提供したりできる可能性があるデータセット内の貴重な情報が失われる可能性がある。したがって、研究者が利用できるデータセットは、セキュリティ上の制約から、情報が不足している可能性がある。複数のデータストアに分散された個人情報または保護された情報を尊重しながら、学習した情報または「知識」を抽出できる技術が、研究者に恩恵を与えることは明らかである。
【０００６】
興味深い点として、分散データの分析に関連する以前の取り組みは、孤立したプライベートデータの技術的な問題を扱うのではなく、機械学習の性質に注目している。例えば、Ｃｏｌｌｉｎｓらにより２００６年１０月２６日に出願された米国特許第７，８９９，２

２５号、発明の名称「Systems and Methods of Clinical State Prediction Utilizing Medical Image Data」は、統計モデルを作成および併合して、最終的な多次元分類空間を作成することを開示している。統計モデルは、被験者が表現される空間を定義する数学的変動モデルである。しかしながら、Ｃｏｌｌｉｎｓでは、予測モデルを構築するために、システムが全てのデータにアクセスする権限を有すると仮定している。また、Ｃｏｌｌｉｎｓは、一元管理されていないデータの安全性を保ちまたは非公開にする必要がある状況について考察していない。この場合も、トレーニング済みモデルを何らかの手法で組み合わせることができれば、有用である。

【０００７】

Ｃｒｉｍｉｎｉｓｉらにより２０１２年６月２１日に出願された米国特許第８，９５４，３６５号、発明の名称「Density Estimation and/or Manifold Learning」について検討する。Ｃｒｉｍｉｎｉｓｉは、モデルを組み合わせる手法に注目するのではなく、データセットを単純化することに注目している。Ｃｒｉｍｉｎｉｓｉは、ラベル付けされていないデータ点間の相対距離または他の関係を維持しながら、ラベル付けされていないデータをより低い次元の空間にマッピングする次元削減技術を開示している。このような技術は、計算量を低減するには有用であるが、異種のプライベートデータセットに依存するモデルをどのように組み合わせるかについては、言及されていない。

【０００８】

データの匿名化に取り組むことを試みる更に別の例として、Ｈｉｇｇｉｎｓらによって２０１４年１月１５日に出願された米国特許出願公開第２０１４／０２２２３４９号、発明の名称「System and Methods for Pharmacogenomic Classification」がある。Ｈｉｇｇｉｎｓは、匿名化された吸収、分布、代謝、および排泄（absorption, distribution, metabolism, and excretion：ＡＤＭＥ）薬物データ内の薬理ゲノミクス集団におけるクラスタを表す代理表現型（surrogate phenotypes）を使用することを記載している。代理表現型は、生の患者データの分類に使用できる学習機械（例えば、サポートベクターマシン）をトレーニングするために使用される。Ｈｉｇｇｉｎｓは、代理表現型に基づいてトレーニング済み学習機械を構築することを提供するが、Ｈｉｇｇｉｎｓでは、最初のトレーニングセットを構築するために匿名化されたデータへのアクセスが必要である。上述のように、データの匿名化によって、トレーニングデータセットの価値の一部が損なわれる。

【０００９】

プライベートデータを格納しているエンティティが多数存在する可能性がある分散環境では、大量の高品質の匿名化されたデータへのアクセスを保証することは、不可能である。これは、新しい学習タスクが開始され、新しいタスクを処理できるデータがまだ存在しない場合に特に当てはまる。したがって、トレーニングが開始される前にデータを匿名化する必要なしに、分散環境においてプライベートデータセットから学習した情報または知識を集約できる学習システムが強く求められている。

【００１０】

本明細書において言及した全ての文献は、引用によって本願に援用されるものとし、これは、個々の文献または特許出願について、特別に、個別に、引用によって本願に援用されると示した場合と同様である。援用される文献の用語の定義または用法が本出願の用語の定義と矛盾しまたは相容れない場合、本出願の用語の定義が優先され、文献における用語の定義は、適用されない。

【００１１】

幾つかの実施形態では、本発明の主題の特定の実施形態を説明しおよび特許請求するために使用する成分の量、濃度等の特性、反応条件等を表す数値は、「約」が付された概ねの数であると解釈される場合がある。すなわち、幾つかの実施形態では、詳細な説明および特許請求の範囲に示す数値パラメータは、特定の実施形態によって実現することが求められる所望の特性に応じて変更できる近似値である。幾つかの実施形態では、数値パラメータは、報告された有効桁数を考慮して、通常の丸め演算を適用したものと解釈される。

なお、本発明の幾つかの広範囲な実施形態における数値範囲およびパラメータは、概数であるが、特定の具体例では、実用的な程度に詳しい数値を示す。本発明の幾つかの実施形態に示す数値は、それぞれの検査測定における標準偏差から生じる必然的な誤差を含むことがある。

【００１２】

文脈に矛盾しない限り、本明細書に記載する全ての範囲は、これらの終点を含むと解釈され、無制限の範囲は、商業的に実用的な値のみを含むと解釈される。同様に、文脈に矛盾しない限り、全ての値のリストは、中間値を含むと見なされる。

【００１３】

以下の詳細な説明および特許請求の範囲内の名詞（英語では不定冠詞および定冠詞が付された単語）は、文脈が特に指定しない限り、複数の意味も包含するものとする。また、以下の記述において、「～の中（in）」の意味は、文脈が特に指定しない限り、「～の中（in）」および「～の上（on）」の両方の意味を包含するものとする。

【００１４】

本明細書における値の範囲の列挙は、単に、この範囲内に含まれる個々の値を個別に言及することに代わる便宜的な表現である。個々の値は、特に指定する場合を除き、個別に指定した場合と同様に、本明細書に組み込まれる。ここに説明する全ての方法は、特に指定しない限り、または文脈によって特に制約されない限り、適切な如何なる順序で実行してもよい。何らかのまたは全ての代表例または例示を示す用語（例えば、「等（such as）」）は、ここに示す実施形態が本発明をより明瞭にすることを意図しているものに過ぎず、特許請求される発明の範囲を制約するものではないことを示している。明細書内の表現によって、特許請求されていない要素が本発明の実施に不可欠であると解釈されることはない。

【００１５】

ここに開示する本発明の主題の代替要素または実施形態のグループは、本発明を限定するものではない。各グループの要素は、個別に特許請求してもよく、他のグループの要素またはここに示す他の要素と組み合わせて特許請求することもできる。利便性および／または特許性の理由に基づいて、あるグループに１つ以上の要素を追加してもよく、あるグループから１つ以上の要素を省略してもよい。このような追加または省略を行った場合、本明細書は、特許請求の範囲で使用される全てのマーカッシュグループの記載を満たすように修正されたグループを含むと見なされる。

【発明の概要】

【発明が解決しようとする課題】

【００１６】

本発明の主題は、分散型オンライン機械学習コンピュータがプライベートデータから情報を学習しまたは知識を獲得し、そのプライベートデータにアクセスできないピア間でその知識を分散させ、分散される知識は、ローカルプライベートデータの実際のプライベートなまたは制限された特徴を含まない装置、システム、および方法を提供することである。

【課題を解決するための手段】

【００１７】

本出願の目的のために、「機械学習」という用語は、明示的にプログラムされることなくデータから学習するように構成された人工知能システムを指すと理解される。このようなシステムは、必然的にコンピュータ技術に根ざしており、実際には、コンピュータ技術がなければ実装できずまたは存在できないと理解される。機械学習システムは、様々なタイプの統計分析を利用するが、機械学習システムは、明示的なプログラミングなしで学習する能力およびコンピュータ技術に根ざしているという点で、統計分析とは、区別される。すなわち、本技術は、学習可能性を保持しながら、プライバシ権を保護する分散データ構造を利用する。生データではなく、圧縮／学習されたデータを交換するプロトコルにより、帯域幅のオーバーヘッドが削減される。

【００１８】
本発明の主題の一側面は、分散型機械学習システムを含む。幾つかの実施形態において、分散型機械学習システムは、分散型演算環境においてピアとして動作する可能性がある複数のプライベートデータサーバを有する。各プライベートデータサーバは、それぞれのローカルプライベートデータにアクセスできる。システム内の他のサーバまたはピアは、通常、他のローカルプライベートデータについての許可、権限、特権、またはアクセスを有していない。更に、各プライベートデータサーバは、例えば、集中型の機械学習コンピュータファームまたは別のプライベートデータサーバ等、グローバルモデリングエンジンを備える１つまたは複数の非プライベート演算デバイスに通信可能に接続されている。プライベートデータサーバは、非一時的コンピュータ可読メモリに格納されたソフトウェア命令を実行するように構成可能な１つまたは複数のプロセッサを有する演算デバイスであり、ソフトウェア命令の実行により、プライベートデータサーバ上にモデリングエンジンが生成される。モデリングエンジンは、ローカルプライベートデータに基づいて、１つまたは複数のトレーニング済み機械学習モデルを生成するように構成可能である。より具体的には、モデリングエンジンは、ネットワークを介して１つまたは複数のリモート演算デバイスからモデル指示を受け取ることができる。モデル指示は、モデリングエンジンに対して、ローカルプライベートデータの少なくとも一部を使用して、機械学習アルゴリズム（例えば、サポートベクターマシン、ニューラルネットワーク、決定木、ランダムフォレスト、深層学習ニューラルネットワーク等）の実装に基づいてトレーニング済み実モデルを作成するように指示する１つまたは複数のコマンドと見なすことができる。モデリングエンジンは、必要であれば、必要な前処理要件（例えば、フィルタリング、検証、正規化等）を実行したのちに、ローカルプライベートデータ（すなわち、選択またはフィルタリングされたトレーニングデータセット）の関数としてトレーニング済み実モデルを作成する。一旦トレーニングされると、トレーニング済み実モデルは、トレーニング済み実モデルの性質を記述する１つ以上の実モデルパラメータまたはメトリクス（例えば、精度、精度ゲイン、感度、感度ゲイン、パフォーマンスメトリクス、重み、学習率、エポック、カーネル、ノード数、層数等）を有することになる。更に、モデリングエンジンはローカルプライベートデータトレーニングセットから１つまたは複数のプライベートデータ分布を生成し、ここで、プライベートデータ分布は、学習モデルの作成に使用されたローカルプライベートデータの性質を表す。モデリングエンジンは、プライベートデータ分布を使用して、プロキシデータのセットを生成し、これは、ローカルプライベートデータと同じ一般的なデータ分布特徴を有しながら、ローカルプライベートデータの実際のプライベートなまたは制限された特徴を含まない合成データまたはモンテカルロデータと見なすことができる。幾つかの場合、擬似乱数生成器のシードを使用して、モンテカルロシミュレーションによって決定論的なプロキシデータのセットを生成する。真の乱数のシードのソースは、例えば、ｒａｎｄｏｍ．ｏｒｇによって提供されている（ＵＲＬ：www.random.orgを参照）。ローカルプライベートデータのプライベートなまたは制限された特徴は、以下に限定されるものではないが、社会保障番号、患者名、住所、またはその他の個人を特定する情報、特にＨＩＰＡＡ法の下で保護されている情報を含む。次に、モデリングエンジンは、プロキシデータのセットからトレーニング済みプロキシモデルを作成することによって、プロキシデータのセットがローカルプライベートデータの代わりになる妥当なトレーニングセットであることの検証を試みる。結果として得られるトレーニング済みプロキシモデルは、実モデルパラメータと同じ属性空間に従って定義された１つまたは複数のプロキシモデルパラメータによって記述される。モデリングエンジンは、プロキシモデルパラメータおよび実モデルパラメータの関数として、トレーニング済み実モデルおよびプロキシモデルが互いにどれほど類似しているかを示す類似性スコアを算出する。モデリングエンジンは、この類似性スコアに基づいて、トレーニング済みモデルに関連する１つまたは複数の情報を送信でき、この情報は、プロキシデータ、実モデルパラメータ、プロキシモデルパラメータ、またはその他の特徴を再現するのに十分なプロキシデータまたは情報のセットを含むことができる。例えば、モデル類似性が（例えば、閾値と比較して）類似性

要件を満たす場合、モデリングエンジンは、プロキシデータのセットを非プライベート演算デバイスに送信でき、非プライベート演算デバイスは、プロキシデータを集約モデルに統合する。

【００１９】

本発明の主題の別の側面は、コンピュータにより実現され、プライベートデータを尊重する分散機械学習の方法を含む。この方法の一実施形態は、プライベートデータサーバが、ローカルプライベートデータの少なくとも一部に基づいてトレーニング済み実モデルを作成するためのモデル指示を受信することを含む。モデル指示は、例えば、機械学習アルゴリズムの実装からトレーニング済み実モデルを構築する要求を含むことができる。プライベートデータサーバ上で実行される機械学習エンジンは、機械学習アルゴリズムの実装を関連するローカルプライベートデータでトレーニングすることによってモデル指示に従ってトレーニング済み実モデルを作成することによって継続できる。結果として得られるトレーニング済みモデルは、トレーニング済みモデルの性質を記述する１つまたは複数の実モデルパラメータを含む。この方法の別のステップは、関連するローカルプライベートデータの性質を記述する１つまたは複数のプライベートデータ分布を生成することを含む。例えば、プライベートデータ分布は、ガウス分布、ポアソン分布、ヒストグラム、確率分布、または他のタイプの分布によって表すことができる。機械学習エンジンは、プライベートデータ分布から、プライベートデータ分布の性質を記述する１つまたは複数の顕著なプライベートデータ特徴（salient private data features）を識別または他の手法で算出できる。分布のタイプに応じて、例示的な特徴は、サンプルデータ、平均、最頻値、代表値、幅、半減期、勾配、モーメント、ヒストグラム、高次モーメント、または他のタイプの特徴を含むことができる。幾つかのより具体的な実施形態では、顕著なプライベートデータ特徴は、プロキシデータを含むことができる。顕著な特徴が利用可能になると、機械学習エンジンは、ネットワークを介して、非プライベート演算デバイス（例えば、主要なプライベートデータ特徴を他のデータセットと統合して集約モデルを作成できる中央サーバやグローバルモデリングエンジン等）に対して、顕著なプライベートデータ特徴を送信する。したがって、複数のプライベートピアが、自らのプライベートデータを公開することなく、学習した知識を共有できる。

【００２０】

本発明の主題の様々な目的、特徴、側面、および利点は、共通の構成要素に共通の参照符号を付した添付の図面とともに、以下の好ましい実施形態の詳細な説明からからより明らかになる。

【図面の簡単な説明】

【００２１】

【図１】ここに提示する実施形態に基づく例示的な分散型オンライン機械学習システムを示す図である。

【００２２】

【図２】ここに提示する実施形態に基づくプライベートデータサーバ内に展開された例示的な機械学習モデリングエンジンアーキテクチャを示す図である。

【００２３】

【図３】ここに提示する実施形態に基づくトレーニング済みプロキシモデルを構築するための準備におけるプロキシトレーニングデータの生成を示すフローチャートである。

【００２４】

【図４】ここに提示する実施形態に基づくトレーニング済み実モデルとトレーニング済みプロキシモデルとの類似性を比較する１つまたは複数の類似性スコアの生成を示すフローチャートである。

【００２５】

【図５】ここに提示する実施形態に基づき、プライベートデータサーバが、実データによって生成されたトレーニング済み実モデルの性質を複製できるプロキシデータを生成し、プロキシデータが非プライベート演算デバイスに送信される、分散型オンライン機械学習

の例示的な方法を示す動作フローチャートである。
【００２６】
【図６】ここに提示する実施形態に基づき、プライベートデータサーバが集約されたプライベートデータの顕著な特徴を非プライベート演算デバイスに送信し、非プライベート演算デバイスがトレーニング済みグローバルモデルに統合するためのプロキシデータを作成する分散型オンライン機械学習の例示的な方法を示す動作フローチャートである。
【発明を実施するための形態】
【００２７】
コンピュータまたは演算デバイスといった文言は、サーバ、インタフェース、システム、機器、データベース、エージェント、ピア、エンジン、コントローラ、モジュール、または個別に、集合的にまたは共同的に動作する他のタイプの演算デバイスを含む、演算デバイスの任意の適切な組み合わせを含むと解釈される。演算デバイスが有形の非一時的コンピュータ可読記憶媒体（例えば、ハードドライブ、ＦＰＧＡ、ＰＬＡ、ＰＬＤ、ソリッドステートドライブ、ＲＡＭ、フラッシュ、ＲＯＭ、外部ドライブ、メモリースティック等）に格納されるソフトウェア命令を実行するように構成された１つまたは複数のプロセッサを備えることは、当業者にとって明らかである。ここに開示する装置に関して以下に説明するように、ソフトウェア命令は、役割、責任、または他の機能を提供するように演算デバイスを具体的に構成またはプログラムする。更に、ここに開示する技術は、コンピュータベースのアルゴリズム、プロセス、方法、または他の命令の実装に関連する開示されたステップまたは動作を実行するためにプロセッサによって実行可能なソフトウェア命令を格納する有形の非一時的コンピュータ可読媒体を含むコンピュータプログラム製品として具現化できる。幾つかの実施形態では、様々なサーバ、システム、データベース、またはインタフェースは、例えば、ＨＴＴＰ、ＨＴＴＰＳ、ＡＥＳ、公開鍵と秘密鍵の交換、ＷｅｂサービスＡＰＩ、既知の金融取引プロトコル、または他の電子情報交換法に基づいて、標準化されたプロトコルまたはアルゴリズムを使用してデータを交換する。デバイス間のデータ交換は、パケット交換ネットワーク、インターネット、ＬＡＮ、ＷＡＮ、ＶＰＮ、または他のタイプのパケット交換ネットワーク、回線交換ネットワーク、セル交換ネットワーク、または他のタイプのネットワークを介して行うことができる。
【００２８】
本明細書および特許請求の範囲では、システム、エンジン、サーバ、デバイス、モジュール、または他の演算要素が、メモリ内のデータに対して機能を実現または実行するように構成されると記述するが、「構成される」または「プログラムされる」という表現の意味は、演算要素のメモリに記憶されたソフトウェア命令のセットによって、演算要素の１つまたは複数のプロセッサまたはコアが、メモリに格納されているターゲットデータまたはデータオブジェクトに対して一連の機能を実行するようにプログラムされることを意味する。
【００２９】
ここに開示する技術は、基礎となる生データのデータプライバシを尊重しながら機械学習データを交換するためのネットワークを介した演算デバイス間の通信チャネルの構築を含む多くの有利な技術的効果を提供する。演算デバイスは、プライバシを含まずに「学習した」情報または知識を互いに交換できる。より具体的には、ここに開示するプライベートデータサーバは、プライベートまたはセキュリティ保護されたデータをリモート演算デバイスに送信するのではなく、１つまたは複数の機械学習アルゴリズムのコンピュータベースの実装を介して、ローカルプライベートデータに関する情報を自動的に「学習」しようと試みる。そして、学習された情報は、プライベートデータにアクセスする権限がない他のコンピュータと交換される。更に、技術的効果は、分散されたプライベートデータおよびこれらの対応するデータ分布からトレーニング済みプロキシモデルを計算によって構築することを含む。
【００３０】
ここに開示する本発明の主題の焦点は、人間の能力を超えて、膨大な量のデジタルデー

タに対して動作できる演算デバイスを構築または構成することである。デジタルデータは、典型的には、患者データの様々な側面を表すが、デジタルデータは、「患者」自体ではなく、患者の１つまたは複数のデジタルモデルの表現である。演算デバイスのメモリ内にこのようなデジタルモデルをインスタンス化することによって、演算デバイスは、演算デバイスのユーザにとって有用な方式でデジタルデータまたはモデルを管理でき、特に分散型オンライン機械学習システムでは、このようなツールがなければ、ユーザは、これらを利用できない。したがって、本発明の主題は、演算デバイスがプライベートデータにアクセスできない環境において、分散型機械学習を改善または最適化することである。

【００３１】

以下の説明は、本発明の主題の多くの例示的な実施形態を提供する。各実施形態は、本発明の要素の単一の組み合わせを表すが、本発明の主題は、開示された要素の全ての可能な組み合わせを含む。すなわち、ある実施形態が要素Ａ、Ｂ、およびＣを含み、他の実施形態が要素Ｂ、およびＤを含む場合、本発明の主題は、明示的な記述がなくても、Ａ、Ｂ、Ｃ、またはＤの他のあらゆる組み合わせを含むと見なされる。

【００３２】

文脈において特段の指定がない限り、本明細書で使用する「接続」という用語は、直接接続（互いに接続される２つの要素が互いに接触する。）および間接接続（２つの要素の間に少なくとも１つの追加要素が存在する。）の両方を含むことを意図する。したがって、「に接続された（coupled to）」および「と接続された（coupled with）」という表現は、同義語として使用される。

【００３３】

以下では、健康管理の観点から、より具体的には、癌患者に関連するゲノム配列データからトレーニング済み機械学習モデルを構築することに関連した説明を行う。但し、ここに開示するアーキテクチャは、腫瘍学以外の他の形式の研究にも適合させることができ、例えば、保険データ、金融データ、ソーシャルメディアプロファイルデータ、人材データ、独自の実験データ、ゲームまたはギャンブルデータ、軍事データ、ネットワークトラフィックデータ、ショッピングまたはマーケティングデータ、または他のタイプのデータ等、生データを保護すべきであるまたはプライベートなものであると考えられる全ての場合に利用できる。

【００３４】

例えば、ここに開示する技術は、「サービスとしての学習」ビジネスモデルの一部として使用できる。このタイプのモデルでは、プライベートデータ（例えば、健康管理データ、ゲノムデータ、企業データ等）を有する組織は、機械学習モデル（例えば、トレーニング済み実モデル、トレーニング済みプロキシモデル等）および他の学習情報を生成することがあり、更に、他のグループ（例えば、ベンチャー企業、他の機関、他の事業等）がこれらのモデルを使用して、これらのデータを分析しまたはローカルデータを有料で研究することを許可することがある。例えば、医療現場では、特定の医療機関の患者から収集されたデータを分析して、機械学習を使用してトレーニング済み実モデルおよび／またはトレーニング済みプロキシモデルを作成できる。異なる医療機関または企業の研究者、データアナリスト、または他の事業者は、料金（例えば、一回限りの料金、定期契約料）を支払ってモデルにアクセスし、例えば、自らのデータを分析しまたはローカルデータを研究することができる。したがって、この例では、機械学習モデルは、システム１００に関する内部データに基づいて生成され、システム１００に関する外部データを分類するために使用できる。

【００３５】

更に他の実施形態では、機械学習サービスを提供する組織は、サードパーティによって提供されたデータを分析して、料金を受け取ることができる。ここで、他の医療機関の研究者、データアナリスト、または他の事業者は、有料で、ローカルプライベートデータと同様に個別に分析できまたはローカルプライベートデータと組み合わすことができる形式でデータを提供し、他の学習済み情報と共に機械学習モデル（例えば、トレーニング済み

実モデルまたはトレーニング済みプロキシモデル）を生成し、これを用いて、後にサードパーティから提供される一連のデータを分析できる。したがって、この例では、機械学習モデルは、システム１００に関する外部データに基づいて生成され、システム１００に関する追加の外部データを分類するために使用できる。

【００３６】

このタイプの「サービスとしての学習」モデルを採用できる他の産業用データは、以下に限定されるものではないが、ゲームデータ、軍事データ、ネットワークトラフィック／セキュリティデータ、ソフトウェア実行データ、シミュレーションデータ等を含む。

【００３７】

機械学習アルゴリズムは、観測データに基づく結論を形成するモデルを作成する。トレーニングデータセットは、教師あり学習のために機械学習アルゴリズムに供給される。この場合、入力および既知の出力をトレーニングデータとして提供することによって、機械学習システムは、このトレーニングデータに基づいてモデルを作成できる。すなわち、機械学習アルゴリズムは、入力を出力にマッピングするマッピング関数を生成する。

【００３８】

他の実施形態として、教師なし学習では、データセットが機械学習システムに入力され、機械学習システムは、データポイントのクラスタリングに基づいてデータを分析する。このタイプの分析では、データの分布または構造を反映したモデルを生成するためにデータの基礎となる構造または分布が使用される。このタイプの分析は、類似性を検出する（例えば、２つの画像が同じである）、異常／異常値を特定する、またはデータセット内のパターンを検出するために使用されることが多い。

【００３９】

上述の２つの手法のハイブリッド型である半教師ありモデル（semi-supervised models）は、教師ありモデルと教師なしモデルの両方を利用してデータを分析する。

【００４０】

機械学習モデルは、（既知の出力または回答なしで）入力に基づいて（例えば、分類または回帰を使用して）出力を予測する。予測は、入力をカテゴリにマッピングすること（例えば、画像の特性が存在するか否かを判定するために画像を分析すること）を含むことができる。このタイプの分析では、出力変数は、クラスメンバーシップを識別するクラスラベルの形式をとる。したがって、この手法を用いて、（例えば、画像が指定された特性を含むか否かに基づいて）カテゴリを選択できる。

【００４１】

回帰分析は、回帰直線とその直線を生成するために使用されるデータ点との間の誤差を最小限に抑えることを目的とする。ここで、出力変数は、連続応答を予測するための連続変数（例えば、線）の形式をとる。したがって、回帰分析は、数値データの分析に使用できる。これらの技術は、後により詳細に説明する。なお、回帰分析は、研究課題の要求に応じて１次元以上の関連次元で行われる。

【００４２】

図１は、例示的な分散型機械学習システム１００を示している。システム１００は、研究者が、通常、アクセスするための許可または権限を有していない多くのプライベートなまたは安全なデータソースから、複数の研究者またはデータアナリストがトレーニング済み機械学習モデルを作成することを可能にするコンピュータベースの研究ツールとして構成されている。図示の例では、研究者は、グローバルモデリングエンジン１３６として実行される非プライベート演算デバイス１３０として表される中央機械学習ハブにアクセスする権限を有する。非プライベート演算デバイス１３０は、研究者に分散型機械学習サービスを提供する１つまたは複数のグローバルモデルサーバ（例えば、クラウド、ＳａａＳ、ＰａａＳ、ＩａａＳ、ＬａａＳ、ファーム等）を備えることができる。なお、研究者にとって関心のあるデータは、ネットワーク１１５（例えば、無線ネットワーク、イントラネット、セルラネットワーク、パケット交換ネットワーク、アドホックネットワーク、インターネット、ＷＡＮ、ＶＰＮ、ＬＡＮ、Ｐ２Ｐ等）を介して、１つ以上のエンティティ

１２０Ａ～１２０Ｎに配置されている１つ以上のプライベートデータサーバ１２４Ａ、１２４Ｂ～１２４Ｎ（まとめてプライベートデータサーバ１２４と呼ぶ。）に存在する。ネットワーク１１５は、上述のネットワークの任意の組み合わせを含むことができる。エンティティは、病院１２０Ａ、診療所１２０Ｂ、実験室１２０Ｎ（集合的にエンティティ１２０と呼ぶ。）を含むことができる。各エンティティ１２０は、自らのローカルプライベートデータ１２２Ａ～１２２Ｎ（まとめてプライベートデータ１２２と呼ぶ。）にアクセスし、これらは、ローカルのストレージ設備（例えば、ＲＡＩＤシステム、ファイルサーバ、ＮＡＳ、ＳＡＮ、ネットワークアクセス可能なストレージデバイス、ストレージエリアネットワークデバイス、ローカルコンピュータ可読メモリ、ハードディスクドライブ、光ストレージデバイス、テープドライブ、テープライブラリ、ソリッドステートディスク等）に保管されていることがある。更に、各プライベートデータサーバ１２４は、ＢＡＭサーバ、ＳＡＭサーバ、ＧＡＲサーバ、ＢＡＭＢＡＭサーバ、または臨床オペレーティングシステムサーバのうちの１つまたは複数を含むことができる。プライベートデータサーバ１２４のそれぞれは、自らのローカルプライベートデータ１２２にアクセスし、少なくとも１つのモデリングエンジン１２６を有する。説明のために、プライベートデータサーバ１２０のそれぞれは、ネットワーク１１５を介して、非プライベート演算デバイス１３０に通信可能に接続されているとする。

【００４３】

プライベートデータ１２２の各セットは、対応するエンティティ１２０にとっては、プライベートであると見なされる。すなわち、この条件の下では、他のエンティティ１２０、並びに非プライベート演算デバイス１３０によって提供されるモデリングサービスにアクセスする研究者は、他のプライベートデータ１２２にアクセスするための権利、許可、または他の権限を有さない。更に明確にするために、用語「プライベート」および「非プライベート」は、エンティティと対応するデータセットとの様々なペアの間の関係を記述する相対的な用語とする。例えば、診療所１２０Ｂのプライベートデータサーバ１２４Ｂは、そのローカルプライベートデータ１２２Ｂにアクセスできるが、他のプライベートデータ、例えば、実験室１２０Ｎのプライベートデータ１２２Ｎまたは病院１２０Ａのプライベートデータ１２２Ａには、アクセスできない。他の実施形態では、プライベートデータサーバ１２４Ｎは、他のエンティティに対する非プライベート演算デバイス１３０と見なすことができる。このような検討は、様々なプライベートサーバ１２４が、中央ハブを介するのではなく、ピアツーピア方式でまたは所属（affiliation）により、ネットワーク１１５を介して互いに直接通信できる実施形態において特に重要である。例えば、医療機関が複数の場所および／または所属機関、例えば、一次医療機関（main hospital）、病院（physician offices）、診療所（clinics）、二次医療機関（secondary hospital）、病院所属機関（hospital affiliation）を有する場合、これらのエンティティのそれぞれが自らのプライベートデータ１２２、プライベートデータサーバ１２４、およびモデリングエンジン１２６を有することができ、これらは全て互いにとって可視であるが、異なるエンティティからは不可視である。

【００４４】

各エンティティ１２０がそのプライベートデータ１２２を安全に維持しなくてはならないというシステムの性質および要件を考慮すると、研究者が、望ましいトレーニング済み機械学習モデルを構築するために必要な大量の高品質データへのアクセス権を獲得することは、困難な課題である。より具体的には、研究者は、関心のあるプライベートデータ１２２を有する各エンティティ１２０から許可を獲得する必要がある。更に、様々な制限（例えば、プライバシポリシー、規則、ＨＩＰＡＡコンプライアンス等）のために、各エンティティ１２０は、要求されたデータを研究者に提供することを許可できない場合がある。研究者がこれらの関連するプライベートデータ１２２を取得するための許可をエンティティ１２０の全てから獲得できると仮定しても、エンティティ１２０は、依然としてデータセットを匿名化する必要がある。このような匿名化は、データを匿名化するのに必要な時間および情報の喪失のために問題となる可能性があり、これは、研究者がトレーニング

機械学習モデルから得る知識の有用性に影響する。

【００４５】

図１に示すエコシステム／システムでは、プライベートデータ１２２のプライバシ制限に関連する問題は、生データ自体ではなく、トレーニング済み機械学習アルゴリズムから得られる知識に注目することによって解決される。研究者は、エンティティ１２０のそれぞれからの生データを要求するのではなく、自らが作成を望む機械学習モデルを定義できる。研究者は、非プライベート演算デバイス１３０を介して、研究者がプライベートデータサーバへのアクセスを許可されていることを条件として、プライベートデータサーバ１２４のうちの１つを介して、または非プライベート演算デバイス１３０とインタフェースできるシステム１００の外部のデバイスを介して、システム１００とインタフェースできる。そして、どのようにして所望のモデルを作成するかについてのプログラムモデル指示が、対応するモデリングエンジン１２６（すなわち、１２６Ａ～１２６Ｎ）を有する各関連プライベートデータサーバ１２４に提出される。各ローカルモデリングエンジン１２６は、自らのローカルプライベートデータ１２２にアクセスし、研究者によって作成されたモデル指示に従ってローカルのトレーニング済みモデルを作成する。各モデリングエンジン１２６が新しい学習情報を獲得し、送信基準が満たされると、新しい知識が非プライベート演算デバイス１３０の研究者に返送される。そして、新しい知識は、グローバルモデリングエンジン１３６を介して、トレーニング済みグローバルモデルに集約できる。知識の例は、以下に限定されるものではないが、プロキシデータ２６０、トレーニング済み実モデル２４０、トレーニング済みプロキシモデル２７０、プロキシモデルパラメータ、モデル類似性スコア、または匿名化された他のタイプのデータを含む（例えば、図２参照）。幾つかの実施形態では、グローバルモデルサーバ１３０は、プロキシ関連情報（例えば、プロキシデータ２６０、プロキシデータ分布３６２、プロキシモデルパラメータ４７５、シードと組み合わせた他のプロキシ関連データ等を含む。）のセットを分析し、このような情報を組み合わせる前に、プライベートデータサーバ１２４のうちの１つからのプロキシ関連情報が別のプライベートデータサーバ１２４からのプロキシ関連データと同じ形状および／または全体的特性を有するか否かを判定する。手動レビューのために、異なるプロキシ関連情報にフラグを立て、基礎となるプライベートデータ分布セットが破損しているか、データが欠落しているか、または相当数の外れ値が含まれているかを判定してもよい。幾つかの実施形態では、外れ値であると見なされるプライベート患者データは、無視され、ここに開示する技術から除外される。例えば、１クラスサポートベクターマシン（support vector machine：ＳＶＭ）を使用して、コアの関連データと一致しない可能性がある外れ値を特定できる。幾つかの実施形態では、外部のピア（例えば、非プライベート演算デバイス１３０等）が、同様の関心データに基づいて、１クラスＳＶＭ（one-class SVM）を構築する。次に、１クラスＳＶＭをプライベートデータサーバ１２４に送信してもよい。そして、プライベートデータサーバ１２４は、外部で生成された１クラスＳＶＭを使用して、関心のあるローカルデータと、実際に関心のある外部データとを確実に一致させることができる。

【００４６】

このように、プロキシデータは、生データの特徴を保持する異なる形式のデータに生データを変換したデータと見なすことができる。

【００４７】

プライベートデータサーバ１２４は、例えば、テスト結果が利用可能になったとき、新しい診断が行われたとき、新しい患者がシステムに追加されたとき等において、新しいプライベートデータ１２２に継続的にアクセス可能である。データセットが比較的小さい場合、プロキシデータ２６０または他のプロキシ関連情報は、格納されているプライベートデータの全部または略全部を使用して再生成できる。データセットがより大きい場合、新しく追加されたデータのみを使用してプロキシデータを再生成してもよい。新しいデータは、タイムスタンプ、格納場所、ジオスタンプ、ブロックチェーンハッシュ等によって識別できる。

【００４８】
他の実施形態では、新しいプライベートデータがリアルタイムまたは略リアルタイムで機械学習システムに組み込まれる。したがって、新しいプライベートデータが利用可能になると、これをトレーニング済み実モデルおよびトレーニング済みプロキシモデルに直ちに組み込むことができる。幾つかの実施形態では、機械学習モデルは、例えば、全ての利用可能なプライベートデータ（古いプライベートデータおよび新しく追加されたプライベートデータ）を使用して、または新しく追加されたプライベートデータのみについて、常時更新される。更に、機械学習モデルの更新を管理する時間枠が設定されていないため、特定の機械学習モデルは、毎日更新され、他のモデルは、毎年更新され、または更に長い時間枠で更新される。この柔軟性は、全てのデータの一括処理とこれに続くトレーニングとテストのサイクルに依存する従来の機械学習モデルとは対照的である。

【００４９】
幾つかの実施形態では、各プライベートデータサーバ１２４は、所望のモデルをどのように作成するかについての同じプログラムモデル指示２３０を受信する。他の実施形態では、あるプライベートデータサーバは、第１のモデルを作成するための第１のプログラムモデル指示セットを受信し、別のプライベートデータサーバは、第２のモデルを作成するための第２のプログラムモデル指示セットを受信する。すなわち、各プライベートデータサーバ１２４に供給されるプログラムモデル指示は、同じであっても異なっていてもよい。

【００５０】
プロキシデータ２６０が生成され、グローバルモデルサーバ１３０に中継されると、グローバルモデルサーバは、データを集約し、更新されたグローバルモデルを生成する。グローバルモデルが更新されると、更新されたグローバルモデルが前のバージョンのグローバルモデルに対して向上しているか否かを判定できる。更新されたグローバルモデルが向上している（例えば、予測精度が向上している）場合、更新されたモデル指示２３０を介して新しいパラメータをプライベートデータサーバに提供してもよい。プライベートデータサーバ１２４では、トレーニング済み実モデルの性能（例えば、モデルが向上するかまたは悪化するか）を評価して、更新されたグローバルモデルによって提供されるモデル指示が、トレーニング済み実モデルを向上させるか否かを判定できる。必要に応じて、以前の機械学習モデルを後に検索できるように、様々な機械学習モデルのバージョンに関連するパラメータを保存してもよい。

【００５１】
更に他の実施形態では、プライベートデータサーバ１２４は、ピアプライベートデータサーバ（異なるプライベートデータサーバ１２４）から、プロキシ関連情報（例えば、プロキシデータ２６０、プロキシデータ分布３６２、プロキシモデルパラメータ４７５、シードと組み合わせた他のプロキシ関連データ等を含む。）を受信してもよい。プライベートデータサーバは、自らのローカルプライベートデータに基づいてまたは自らのローカルプライベートデータとピアプライベートデータサーバから受信したプロキシ関連情報との両方に基づいて、モデルを生成できる。組み合わされたデータセットの予測精度が向上する場合、データセットまたは学習された知識が組み合わされる。

【００５２】
幾つかの実施形態において、情報（例えば、トレーニング済みプロキシモデルを含む機械学習モデル、トレーニング済み実モデル、プライベートデータ分布、合成／プロキシデータ分布、実モデルパラメータ、プロキシモデルパラメータ、類似性スコア、または機械学習プロセスの一部として生成されたその他の情報等）に（処理が発生した場所に関連付けられまたは場所を示す他の識別子に関連付けられている）ジオスタンプを付してもよく、タイムスタンプを付してもよく、ブロックチェーンに統合して調査をアーカイブしてもよい（ＵＳ２０１５０３３２２８３参照）。ブロックチェーンは、サンプル固有の監査証跡（audit trails）として構成できる。この例では、ブロックチェーンは、単一のサンプルの単一の独立したチェーンとしてインスタンス化され、サンプルのライフサイクルまた

は監査証跡を表す。更に、システムが非同期的に新しいデータを継続的に受信できるため、ジオスタンプは、新しい情報の流入を管理するのに役立つ（例えば、新しく追加された診療所については、その診療所からのものとしてジオスタンプされた全てのデータが機械学習システムに組み込まれる）。如何なるタイプのデータにジオスタンプを付してもよい。

【００５３】

図２は、機械学習活動に関してエンティティ２２０内にプライベートデータサーバ２２４を含む例示的アーキテクチャを示している。図２に示す例は、プライベートデータサーバ２２４がリモート演算デバイスおよびプライベートデータ２２２とどのようにインタラクトするかという観点から見た本発明の概念を示す。より好ましい実施形態では、プライベートデータ２２２は、ローカルのプライベート健康管理データを含み、またはより具体的には、患者固有のデータ（例えば、氏名、ＳＳＮ、正常ＷＧＳ、腫瘍ＷＧＳ、ゲノム差分オブジェクト（genomic diff objects）、患者識別子等）を含む。エンティティ２２０は、通常、プライベートなローカルの生データを有し、上述したような制限を受ける機関である。エンティティの例としては、病院、研究所、診療所、薬局、保険会社、腫瘍専門医のオフィス、またはローカルに保存されたデータを有する他のエンティティが含まれる。プライベートデータサーバ２２４は、通常、エンティティ２２０のファイアウォールの背後に配置されたローカルサーバを表す。プライベートデータサーバ２２４は、メモリ２９０に格納されているソフトウェア命令２９３を実行するように構成されている１つまたは複数のプロセッサ２９７を有するコンピュータとして具現化できる。本発明の主題に利用できる例示的なサーバは、Ｌｉｎｕｘ（登録商標）サーバ、Ｗｉｎｄｏｗｓ（登録商標）サーバ、または他のサーバを含む。

【００５４】

プライベートデータサーバ２２４は、エンティティ２２０の関係者（stakeholders）に代わってプライベートデータ２２２へのアクセスを提供する。より好ましい実施形態では、プライベートデータサーバ２２４は、特定の患者データ、特に、大きなサイズのデータセットのローカルキャッシュを表す。例えば、患者は、癌のための様々な治療を受けている場合もあり、あるいは、臨床試験に参加している場合もある。このようなシナリオでは、患者のデータは、１つ以上のゲノム配列データセットを含むことができ、各データセットは、数百ギガバイトのデータを含むことができる。複数の患者がいる場合、合計データセットは、数テラバイト以上である可能性がある。例示的なゲノム配列データセットは、全ゲノム配列（whole genome sequence：ＷＧＳ）、ＲＮＡ配列データ、全エキソーム配列（whole exome sequence：ＷＥＳ）、プロテオミクスデータ、組織間の差異（例えば、罹患組織ｖｓ対応正常組織、腫瘍組織ｖｓ対応正常組織、一患者ｖｓ他患者等）、または他の大きなデータセットを含むことができる。更に、患者は、ファイル上に複数のゲノム配列データセット、すなわち腫瘍ＷＧＳと対応する正常ＷＧＳを有することがある。特に興味深い１つのデータセットには、腫瘍配列と、対応正常配列とのゲノム差分が含まれ、これは、「ゲノム差分オブジェクト」と呼ばれることもある。このようなゲノム差分オブジェクトおよびこれらの生成は、Ｓａｎｂｏｒｎらによって、いずれも発明の名称「BAMBAM: Parallel comparative Analysis of High Throughput Sequencing Data」として、２０１１年5月２５日および２０１１年１１月１８日にそれぞれ出願された米国特許第９，６５２，５８７号および米国特許第９，６４６，１３４号に十分に開示されている。別のタイプのデータには、患者の試料に由来する推定プロテオーム経路（inferred proteomic pathways）が含まれ、これは、Ｖａｓｋｅらによって、いずれも発明の名称「Pathway Recognition Algorithm Using Data Integration on Genomic Models (Paradigm)」として、２０１
１年４月２９日および２０１１年１０月２６日にそれぞれ出願された米国特許出願公開第２０１２／００４１６８３号および第２０１２／０１５８３９１号に開示されている。

【００５５】

プライベートデータサーバ２２０を介してこのような大きなデータセットのローカルキ

ャッシュを提供することは、複数の理由から有利であると考えられる。このようなデータセットは、サイズが大きく、オンデマンドで、すなわち要求に応じて直ちに入手することが困難である。例えば、５０倍の読み取りによる患者の完全なＷＧＳのデータのサイズは、概ね１５０ＧＢにもなる。患者の腫瘍の同様のＷＧＳと組み合わせると、データセットのサイズは、３００ＧＢを容易に超えてしまう。これは、当然、単一の腫瘍ＷＧＳと単一の正常ＷＧＳしかないとの仮定に基づいている。異なる腫瘍位置または異なる時間に採取された複数のサンプルがある場合、データセットのサイズは、一人の患者について、テラバイトを容易に超える可能性がある。このような大きなデータセットをダウンロードしまたはデータセットにリモートでアクセスするには時間がかかり、患者をリアルタイムで治療する際に必要とされる緊急性に対応できない。したがって、患者データのローカルキャッシュを有することは、患者および他の関係者にとって最良である。更に、患者が移動したりまたは様々なエンティティと関わったりする際にリアルタイムでデータを移動させることも実用的ではない。データセットが大規模でキャッシュ内に収まらない可能性がある場合、キャッシュされたデータを提供する代わりに、プライベートデータを模倣したミニモンテカルロ（mini Monte Carlo）シミュレーションを使用できる。これらのタイプのシミュレーションは、通常、シードを使用し、シードおよび擬似乱数発生器のパラメータに基づいて、モンテカルロシミュレーションによって、決定論的な手法で合成プライベートデータ（synthetic private data）を生成することを可能にする。データの変更を最小限に抑えながら、好ましい量の合成プライベートデータを生成するシードが特定されると、任意のプライベートデータサーバ１２４にこのシードを提供でき、ここで、このシードを用いて、同じ擬似乱数発生器および他のアルゴリズムを使用して、合成プライベートデータが再生成される。合成データを分析することにより、秘匿するべき識別的特徴が含まれていないことを確認できる。

【００５６】

図示の例では、ソフトウェア命令２９３は、モデリングエンジン２２６の能力または機能を実現する。モデリングエンジン２２６は、プライベートデータ２２２を使用して、機械学習アルゴリズム２９５の１つまたは複数の実装をトレーニングする。機械学習アルゴリズムの実装の例示的なソースの例として、ｓｃｉ-ｋｉｔ　ｌｅａｒｎ、ＴｅｎｓｏｒＦｌｏｗ（商標）を含むＧｏｏｇｌｅ（登録商標）の人工知能、ＯｐｅｎＡＩ（商標）、Ｐｒｅｄｉｃｔｉｏｎ　ＩＯ（商標）、Ｓｈｏｇｕｎ（商標）、ＷＥＫＡ、またはＭａｈｏｕｔ（商標）、Ｍａｔｌａｂ、アマゾン（Amazon）社の機械学習、マイクロソフト（Microsoft）社のアジュール（Azure）機械学習、およびＳｃｉＫｉｔ－Ｌｅａｒｎ等がある。モデリングエンジン２２６内に描かれている様々な要素は、モデリングエンジン２２６内のデータと様々な機能モジュールとのインタラクションを表している。すなわち、モデリングエンジン２２６は、プライベートデータ２２２へのインタフェースを提供するように構成されたローカルエージェント、並びに遠隔の研究者がネットワーク２１５を介してモデリングエンジン２２６内でローカルにトレーニング済みモデルを作成できる経路（conduit）と見なすことができる。現実的には、モデリングエンジン２２６は、ローカルプライベートデータ２２２を、外部の演算デバイスによって消費可能なプライバシを含まないデータに関する知識に変換する変換モジュールである。この知識は、匿名化されている機械学習システムによって生成されるあらゆる情報を含むことができる。

【００５７】

プライベートデータサーバ２２４は、多くの異なる形式をとることができる。幾つかの実施形態では、プライベートデータサーバ２２４は、エンティティ２２０のＩＴインフラストラクチャ内に統合された演算機器、例えば、プライベートデータ用の独自のストレージシステム２２２を有する専用サーバである。このような手法は、プライベートデータ２２２が、エンティティ２２０の外部の特定の研究プロジェクトをターゲットとしている大きなデータセットに関連する状況において有利であると考えられる。例えば、この機器は、政府または臨床試験に高い関連性を有する患者データを保存できる。他の実施形態では、プライベートデータサーバ２２４は、エンティティ２２０のＩＴ部門によって所有およ

び運営される１つまたは複数のサーバを含むことができ、このサーバは、エンティティ２２０のサーバに配備できる追加のソフトウェアモデリングエンジンアプリケーションを含む。

【００５８】

ここに示す例では、プライベートデータサーバ２２４は、ネットワーク２１５を介して通信するように構成可能な演算デバイスとして示されている。説明のため、ネットワーク２１５は、インターネットであるとする。但し、ネットワーク２１５は、ＶＰＮ、イントラネット、ＷＡＮ、Ｐ２Ｐネットワーク、携帯ネットワーク、または他の形態のネットワークを含む他の形態のネットワークであってもよい。プライベートデータサーバ２２４は、リモートデバイスとの接続を確立するために１つまたは複数のプロトコルを使用するように構成可能である。このような通信に利用できるプロトコルの例には、ＨＴＴＰ、ＨＴＴＰＳ、ＳＳＬ、ＳＳＨ、ＴＣＰ／ＩＰ、ＵＤＰ／ＩＰ、ＦＴＰ、ＳＣＰ、ＷＳＤＬ、ＳＯＡＰ、または他のタイプの周知のプロトコルが含まれる。なお、このようなプロトコルを利用できるが、エコシステム／システム内のデバイス間で交換されるデータは、演算デバイスによる容易な転送および消費のために、更にパッケージ化してもよい。例えば、システム内で交換される様々なデータ要素（例えば、モデル指示２３０、プロキシデータ２６０等）は、１つまたは複数のマークアップ言語（例えば、ＸＭＬ、ＹＡＭＬ、ＪＳＯＮ等）または他のファイルフォーマット（例えば、ＨＤＦ５等）を介してパッケージ化してもよい。

【００５９】

幾つかの実施形態では、プライベートデータサーバ２２４は、ネットワークセキュリティインフラストラクチャ、例えば、ファイアウォールの背後に配置される。このような場合、適切なネットワークアドレス変換（network address translation：ＮＡＴ）ポートがファイアウォール内に作成されていない限り、リモート演算デバイスは、プライベートデータサーバ２２４との接続を確立できない可能性が高い。しかしながら、より好ましい手法は、例えば、モデリングエンジン２２６を介して、ファイアウォールを越えて、中央モデリングサーバ（例えば、図１の非プライベート演算デバイス１３０）との通信リンクを確立するようにプライベートデータサーバ２２４を構成することである。この手法は、ファイアウォールの変更を必要としないため有利である。この場合も、通信リンクは、暗号化（例えば、ＨＴＴＰＳ、ＳＳＬ、ＳＳＨ、ＡＥＳ等）を介して保護できる。

【００６０】

モデリングエンジン２２６は、プライベートデータサーバ２２４内で動作するエージェントを表し、トレーニング済み機械学習モデルを作成するように構成可能である。幾つかの実施形態では、モデリングエンジン２２６は、特定の研究タスク専用の安全な仮想マシンまたは安全なコンテナ内で機能でき、これによって、各研究者の作業を互いにセキュリティ保護しながら、複数の異種の研究者が並行して作業することが可能になる。例えば、モデリングエンジン２２６は、Ｄｏｃｋｅｒ（登録商標）コンテナを介して実装でき、ここで、各研究者は、プライベートデータサーバ２２４上で動作する自らのモデリングエンジン２２６の個別のインスタンスを有することになる。他の実施形態では、モデリングエンジン２２６は、多数のセッションを並行して処理するように構成でき、各セッションは、プライベートデータサーバ２２４のオペレーティングシステム（例えば、Ｌｉｎｕｘ、Ｗｉｎｄｏｗｓ等）内の個別のスレッドとして実装できる。

【００６１】

プライベートデータサーバ２２４と１つまたは複数のリモート非プライベート演算デバイスとの間に通信リンクが確立されると、モデリングエンジン２２６は、そのサービスを外部のエンティティ、例えば、研究者に提供する準備が整う。モデリングエンジン２２６は、プライベートデータ２２２の少なくとも一部の関数としてトレーニング済み実モデル２４０を作成するようにモデリングエンジン２２６に指示する１つまたは複数のモデル指示２３０を受信する。例えば、ニューラルネットワーク等の幾つかの実施形態では、入力および他の構成パラメータは、モデル指示によって提供され、各入力の重みは、機械学習

システムによって決定される。トレーニング済み実モデル２４０は、機械学習アルゴリズム２９５の実装からトレーニングされたトレーニング済み機械学習モデルである。トレーニングが完了した後、トレーニング済み実モデル２４０は、１つ以上のトレーニング済みモデルパラメータ２４５を含むことになる。

【００６２】

モデリングエンジン２２６は、機械学習アルゴリズム２９５の実装に応じて、ローカルプライベートデータ２２２の少なくとも一部からトレーニング済み実モデル２４０を作成するためのモデル指示を受信する。モデル指示２３０は、プライベートデータ２２２から知識を取得するようにモデリングエンジン２２６を構成できる多くの可能なメカニズムを表し、エンティティ２２０内で生成されたローカルコマンド、ネットワーク２１５を介して供給されたリモートコマンド、実行可能ファイル、プロトコルコマンド、オプションのメニューから選択されたコマンド、または他のタイプの命令を含むことができる。モデル指示２３０は、所望の実装に応じて大きく異なることがある。幾つかの場合、モデル指示２３０は、例えば、スクリプト（例えば、パイソン（Python）、ルビー（Ruby）、ジャバスクリプト（JavaScript（登録商標））等）の形式で、所望のトレーニング済みモデルをどのように作成するかについてモデリングエンジン２２６に通知するストリームライン指示を含むことができる。更に、モデル指示は、プライベートデータ２２２から作成された所望の結果セットに対する要件、並びにどの機械学習アルゴリズム２９５を使用すべきかを定義するデータフィルタまたはデータ選択基準を含むことができる。研究者が、サポートベクターマシン（support vector machine：ＳＶＭ）に基づいて、患者の腫瘍配列と、患者の対応する正常配列との間の特定のゲノムの相違を考慮して、どの患者が様々な薬物に対するレスポンダー（responder）であるか、またはノンレスポンダー（non-responder）であるかを調査するシナリオについて検討する。このような場合のモデル指示２３０は、例えば、ＸＭＬまたはＨＤＦ５を介してパッケージ化された、プライベートデータ２２２から選択されるデータの要件、特定された薬物、特定のゲノム差分オブジェクトへの参照、応答ｖｓ非応答等を含むことができる。モデル指示２３０は、例えば、識別子（例えば、番号、氏名、ＧＵＩＤ等）およびバージョン番号による所望のＳＶＭへの特定の参照を含むことができ、あるいはモデリングエンジン２２６が実行するために準備されたＳＶＭの予めパッケージ化された実装を含んでいてもよい。

【００６３】

幾つかの実施形態では、メタデータを収集するように構成されたアプリケーションは、プライベートデータをスキャンして、プライベートデータリポジトリに格納されているデータのタイプを判定できる。例えば、このアプリケーションは、ファイルリポジトリをスキャンして、存在するファイルのタイプを識別する（例えば、特定のタイプのデータが利用可能であることを示す特定のプログラムに固有のファイル名拡張子を識別する、使用可能なデータのタイプ等を示す命名規則に従って命名されたファイル名をスキャンする、など）ことができる。他の実施形態では、アプリケーションがデータベースと通信して利用可能なデータのタイプをクエリしてもよく、あるいは、利用可能なデータのタイプを反映するレポートをグローバルモデリングサーバ１３０に送信するようにデータベースを構成してもよい。（プライベートデータを反映する）メタデータの記述が利用可能になると、モデル指示は、プライベートデータを参照するように構成でき、これにより、機械学習システムへの入力の選択に関する指示を提供する。研究者によるクエリが継続的および定期的に、例えば、定期的な間隔で更新される場合、システムは、メタデータを認識し、重要なパラメータが存在するか否かを判定し、研究者によって設定されたクエリに対応するモデル指示の生成および送信を引き起こすように構成できる。他のケースでは、新しいクエリに対して、研究者は、手動または半自動の手法でモデル指示を生成してもよい。新しいクエリに対して、システムは、このような新しいクエリに対するモデル指示を生成するために分析するデータのタイプに関する推奨を提供するように構成してもよい。

【００６４】

各プライベートデータサーバからのメタデータは、グローバルモデルサーバに提供でき

る。メタデータは、（生データやプライベートデータではない）属性スペースを返す。この情報に基づいて、機械学習タスクを生成する研究者は、特定のプライベートデータのセットを分析するために特定のプライベートデータサーバに対するモデル指示を構成できる。

【００６５】

幾つかの実施形態では、プライベートデータサーバは、モデルの精度が低いことを認識し、グローバルモデルサーバから追加の更新を要求できる。グローバルモデリングエンジンを使用するグローバルモデルサーバは、様々な場所からのデータをグローバルモデルに集約する。例えば、癌生存モデルの向上が要求され、同じタイプのデータが利用できない場合、癌生存モデルの予測精度を向上するために異なる組織型からのデータを組み合わせてもよい。

【００６６】

モデル指示２３０は、遥かに複雑な性質を有することもできる。具体的には、モデル指示２３０は、自己完結型であってもよく、この場合、実際には、ローカルデータベースとインタフェースされるように特別に構成されたクエリエンジン（例えば、ＳＱＬ、ＮｏＳＱＬ等）、機械学習アルゴリズム２９５の予めコンパイルされた（例えば、オブジェクトコード、バイトコード等）実装、結果として生じるモデルを管理するための規則、等を含む完全なモデリングパッケージを含んでいてもよい。このような手法は、パッケージ化された送達可能なコンテナを介して実施できる。なお、モデル指示２３０は、単純な構成から、上述したより複雑な構成までの範囲内で変更できる点を十分に考慮する必要がある。すなわち、モデル指示２３０は、ローカルコンピュータから受信するローカルコマンド、ネットワーク２１５を介してコンピュータ（例えば、ピアデータサーバまたはグローバルモデルサーバ）から受信するリモートコマンド、実行可能ファイル、プロトコルコマンド、オプションのメニューから選択されるコマンド、リモートプロシージャコール、またはその他のタイプの指示を含むことができる。

【００６７】

モデリングエンジン２２６は、モデル指示２３０からのデータ選択基準を利用して、例えば、プライベートデータ２２２を格納しているデータベースにクエリを提出することによって、プライベートデータ２２２から結果セットを作成する。例えば、クエリは、プライベートデータ２２２に格納されている属性またはテーブルにアクセスしまたはこれを読み出すためにモデル指示２３０内の要件から適切にフォーマットされたＳＱＬクエリを含むことができる。結果セットは、データ選択基準の性質に応じて、プライベートデータ２２２と同じ正確なデータまたは適切なサブセットであってもよい。結果セットは、トレーニング済み実モデル２４０に対するトレーニングデータとなる。すなわち、結果セットは、実モデル２４０をトレーニングするために使用できる。健康管理の文脈では、結果セットは、患者データを含み、患者データは、患者固有の情報、例えば、症状、検査、検査結果、提供者名、患者名、年齢、住所、診断、ＣＰＴコード、ＩＣＤコード、ＤＳＭコード、関係、または患者の説明に利用できるその他の情報のうちの１つまたは複数を含むことができる。なお、機械学習アルゴリズムは、ローカルプライベートデータに対して動作するので、結果セットは、データから秘密情報を除去（サニタイズ：sanitize）するための前処理匿名化ステップ（pre-processing de-identification step）を必要としない。患者固有の情報を保持することにより、モデリングエンジン２２６は、この他の場合に失われる可能性がある知識をトレーニング済み実モデル２４０から得ることができるので、ここに開示する手法は、従来の手法より優れていると考えられる。例えば、モデリングエンジン２２６による分析の前に患者名が削除されると、関連するファミリー履歴を予測パラメータとして実モデル２４０に組み込むことができない可能性がある。

【００６８】

モデリングエンジン２２６は、プライベートデータ２２２の少なくとも一部を表す結果セットの関数として、トレーニング済み実モデル２４０を作成する。これは、モデリングエンジン２２６が機械学習アルゴリズム２９５の所望の実装をプライベートデータ２２２

の結果セットでトレーニングすることによって達成される。所望の機械学習アルゴリズム２９５が多様なアルゴリズムを含むことができることを考慮すると、モデル指示２３０は、トレーニングが行われる条件を定義する指示を含むことができる。例えば、この条件は、トレーニングデータ、学習速度、収束要件、トレーニングの制限時間、初期条件、感度、特異性、または必須の若しくはオプションのその他のタイプの条件について実行するための複数個の繰り返しまたはエポックを含むことができる。収束要件は、「変化率」等の一次導関数、「加速度」等の二次導関数、またはデータの属性空間における他の次元の高次時間導関数または更に高次の導関数等を含むことができる。

【００６９】

機械学習アルゴリズム２９５は、分類アルゴリズム、ニューラルネットワークアルゴリズム、回帰アルゴリズム、決定木アルゴリズム、クラスタリングアルゴリズム、遺伝的アルゴリズム、教師あり学習アルゴリズム、半教師あり学習アルゴリズム、教師なし学習アルゴリズム、深層学習アルゴリズム、または他のタイプのアルゴリズムの実装を含む多数の異なるタイプのアルゴリズムを含むことができる。より具体的には、機械学習アルゴリズム２９５は、サポートベクターマシン、決定木、最近傍アルゴリズム、ランダムフォレスト、リッジ回帰、Ｌａｓｓｏアルゴリズム、ｋ－ｍｅａｎｓクラスタリングアルゴリズム、ブースティングアルゴリズム、スペクトルクラスタリングアルゴリズム、平均シフトクラスタリングアルゴリズム、非負行列因数分解アルゴリズム、エラスティックネットアルゴリズム、ベイズ分類アルゴリズム、ＲＡＮＳＡＣアルゴリズム、直交マッチング追跡アルゴリズム、ブートストラップ集約、時差学習、バックプロパゲーション、オンライン機械学習、Ｑラーニング、確率勾配降下、最小二乗回帰、ロジスティック回帰、通常最小二乗回帰（ordinary least squares regression：ＯＬＳＲ）、線形回帰、段階的回帰、多変量適応回帰スプライン（multivariate adaptive regression splines：ＭＡＲＳ）、局所推定散布図平滑化（locally estimated scatterplot smoothing：ＬＯＥＳＳ）アンサンブル法、クラスタリングアルゴリズム、重心ベースのアルゴリズム、主成分分析（principal component analysis：ＰＣＡ）、特異値分解、独立成分分析、ｋ最近傍（k nearest neighbors：ｋＮＮ）、学習ベクトル量子化（learning vector quantization：ＬＶＱ）、自己組織化マップ（self-organizing map：ＳＯＭ）、局所重み付き学習（locally weighted learning：ＬＷＬ）、アプリオリアルゴリズム、ｅｃｌａｔアルゴリズム、正則化アルゴリズム、リッジ回帰、最小絶対収縮および選択演算子（least absolute shrinkage and selection operator：ＬＡＳＳＯ）、エラスティックネット、分類および回帰木（classification and regression tree：ＣＡＲＴ）、反復２分割器３（iterative dichotomiser 3：ＩＤ３）、Ｃ４．５およびＣ５．０、カイ２乗自動インタラクション検出（chi-squared automatic interaction detection：ＣＨＡＩＤ）、決定株、Ｍ５、条件付き決定木、最小角度回帰（least-angle regression：ＬＡＲＳ）、単純ベイズ、ガウス単純ベイズ、多項ナイーブベイズ、平均単依存推定（averaged one-dependence estimator：ＡＯＤＥ）、ベイジアンビリーフネットワーク（bayesian belief network：ＢＢＮ）、ベイジアンネットワーク（bayesian network：ＢＮ）、ｋメジアン、期待値最大化（expectation maximisation：ＥＭ）、階層的クラスタリング、パーセプトロン逆伝搬、ホップフィールドネットワーク、動径基底関数ネットワーク（radial basis function network：ＲＢＦＮ）、ディープボルツマンマシン（deep boltzmann machine：ＤＢＭ）、ディープビリーフネットワーク（deep belief network：ＤＢＮ）、畳み込みニューラルネットワーク（convolutional neural network：ＣＮＮ）、積層型オートエンコーダ、主成分回帰（principal component regression：ＰＣＲ）、部分最小二乗回帰（partial least squares regression：ＰＬＳＲ）、サモンマッピング、多次元スケーリング（multidimensional scaling：ＭＤＳ）、射影追跡、線形判別分析（linear discriminant analysis：ＬＤＡ）、混合判別分析（mixture discriminant analysis：ＭＤＡ）、２次判別分析（quadratic discriminant analysis：ＱＤＡ）、フレキシブル判別分析（flexible discriminant analysis：ＦＤＡ）、ブートストラップ集計（バギング）、アダブースト（adaboost）、積み上げ一般化（ブレンディング）、勾配ブーストマシン（gradient boosting ma

chines：ＧＢＭ）、勾配ブースト回帰木（gradient boosted regression tree：ＧＢＲＴ）、ランダムフォレスト、または今後開発されるアルゴリズムのうちの１つまたは複数の実装形態を含むことができる。トレーニングは、教師あり（supervised）、半教師あり（semi-supervised）、または教師なし（unsupervised）であってもよい。幾つかの実施形態では、機械学習システムは、自然言語処理（Natural Language Processing：ＮＰＬ）を使用してデータ（例えば、音声データ、テキストデータ等）を分析してもよい。トレーニング済み実モデル２４０は、トレーニングされると、学習されたことまたは機械学習ジョブを提出する研究者が望むプライベートデータ２２２から取得された知識を表す。トレーニング済み実モデル２４０は、受動モデル（passive model）または能動モデル（active model）と見なすことができる。受動モデルは、これ以上の作業が行われない最終的な完成モデルを表す。能動モデルは、動的であり、様々な状況に基づいて更新できるモデルを表す。幾つかの実施形態では、トレーニング済み実モデル２４０は、毎日、毎週、隔月、毎月、毎四半期、または毎年、リアルタイムで更新される。（例えば、モデル指示２３０の更新、時間の経過、新しいまたは修正されたプライベートデータ２２２等によって）新しい情報が利用可能になると、能動モデルは、更に更新される。このような場合、能動モデルは、その更新に関してモデルの状態を記述するメタデータを搬送する。メタデータは、バージョン番号、更新日、更新に使用された新しいデータの量、モデルパラメータのシフト、収束要件、またはその他の情報のうちの１つ以上を記述する属性を含むことができる。このような情報は、モデルの大規模なコレクションを長期に亘って管理するために提供され、ここで、各能動モデルは、個別の管理可能なオブジェクトとして扱うことができる。

【００７０】

トレーニング済み実モデル２４０は、匿名化されていない実データについてトレーニングされたものであり、これがプライベートデータ２２２から得られた実際のデータであると考えられることを明確にするために、「実（actual）モデル」と呼んでいる。これは、プロキシデータ２６０についてトレーニングされ、シミュレートされたデータと見なされる、トレーニング済みプロキシモデル２７０（後述する）とは対照的である。

【００７１】

トレーニング済み実モデル２４０は、複数の関心点（points of interest）を含む。第１に、図示はしていないが、トレーニング済み実モデル２４０は、トレーニング済みモデルの性質を記述する上述のメタデータを含むことができる。第２に、トレーニング済み実モデル２４０は、実モデルパラメータ２４５によって表される幾つかのパラメータを含む。実モデルパラメータ２４５は、生データを処理する際に、予測目的でトレーニング済み実モデル２４０によって使用される特定の値である。したがって、実モデルパラメータ２４５は、プライベートデータ２２２からトレーニング済み実モデル２４０を作成することによって取得された知識の抽象的な表現と見なすことができる。実モデルパラメータ２４５がパッケージ化され、リモートの非プライベート演算デバイスまたはピアプライベートデータサーバに送信されると、リモートの非プライベートまたはピア演算デバイスは、プライベートデータ２２２へのアクセスを必要とせずに、リモートの演算デバイスにおいてローカルで、パラメータからトレーニング済み実モデル２４０の新しいインスタンスをインスタンス化することによって、トレーニング済み実モデル２４０を正確に再構築でき、したがって、匿名化の必要性がなくなる。実モデルパラメータ２４５は、トレーニング済み実モデル２４０の性質、およびその基礎となる機械学習アルゴリズム２９５の実装、並びに実モデル２４０を生成するために使用されるプライベートデータ２２２の品質に依存する。実モデルパラメータ２４５の例は、重み、カーネル、層、ノード数、感度、精度、精度向上、ハイパーパラメータ、またはトレーニング済み実モデル２４０を再インスタンス化するために活用できる他の情報を含む。

【００７２】

プライベートデータ２２２の量が高品質で十分なサイズであると考えられる幾つかの実施形態では、実モデルパラメータ２４５をリモートデバイスに送信することは、大いに有

益となることがある。しかしながら、エンティティ２２０は、研究タスクを完了するために十分に大量のローカルデータを有していない場合がある。更に、ここに開示する技術に従って対処される別の課題は、トレーニング済み実モデル２４０から取得された知識を他のエンティティ２２０からのデータとどのように統合するか、および特に、研究者や臨床医の関心を反映したモデル指示に関する知識を生成するために、エコシステム内のピア間の知識をどのように集約するかを含む。ここに示す例では、これらの課題は、モデリングエンジン２２６の構成を介して、トレーニング済み実モデル２４０を作成するために使用されるデータの理解を獲得することによって解決される。

【００７３】

図示の例では、モデリングエンジン２２６は、トレーニング済み実モデル２４０を作成するために使用されるトレーニングデータセットを分析し、プライベートデータ分布２５０によって表されるようなトレーニングデータセットの性質の解釈を生成する。すなわち、モデリングエンジン２２６は、トレーニング済み実モデル２４０を作成するためのトレーニングセットとして使用されるローカルプライベートデータを集約的に表す複数のプライベートデータ分布２５０を生成するように更に構成可能である。幾つかの実施形態では、モデリングエンジン２２６は、可能であれば教師なしの手法で、プライベートデータ分布２５０を介して表すことができるデータ内の関係を発見しようとする試みにおいて、トレーニングデータセットに対して多くの異なるアルゴリズム（例えば、回帰、クラスタリング等）を自動的に実行できる。プライベートデータ分布２５０は、プライベートデータトレーニングセットの全体的な性質を記述する。例えば、プライベートデータ分布２５０は、患者の年齢のヒストグラムを含むことができる。プライベートデータ分布２５０に関する詳細については、図３を用いて後述する。プライベートデータ分布２５０は、連続的、非連続的、離散的、または他のタイプの分布であってもよい。プライベートデータ分布は、以下に限定されるものではないが、ベルヌーイ分布、ラデマッハ分布、二項分布、ベータ二項分布、縮退分布、離散一様分布、超幾何分布、およびポアソン二項分布等の分布を含むことができる。更に、プライベートデータの分布は、負のベータ二項分布、ボルツマン（Boltzmann）分布、ギブス（Gibbs）分布、マクスウェルーボルツマン（Maxwell-Boltzmann）分布、ボレル（Borel）分布、チャンパーノウン（Champernowne）分布、拡張された負の二項分布、拡張された超幾何分布、対数級数分布、対数分布、負の二項分布、複合ポアソン分布、パラボラフラクタル分布、ポアソン分布、ポリアーエッゲンベルガー（Polya-Eggenberger）分布、スキュー楕円分布、ユールーサイモン（Yule-Simon）分布、およびゼータ分布を含むことができる。更に、プライベートデータ分布は、逆正弦分布、ベータ分布、対数正規分布、一様分布、アーウィンーホール（Irwin-Hall）分布、ベイツ（Bates）分布、ケント（Kent）分布、対数分布、マルチェンコーパスツール（Marchenko-Pastur）分布、密度分布、二乗余弦分布、逆数分布、三角分布、台形分布、切り捨て正規分布、U二次分布、およびフォンミーゼスーフィッシャー（von Mises-Fisher）分布を含むことができる。更に、分布は、連続一様分布、スケルラム（Skellam）分布、カイ二乗分布、ガンマ分布、または統計学で使用される他の任意の形式の分布を含む。

【００７４】

モデリングエンジン２２６は、プライベートデータ分布２５０に従ってプロキシデータ２６０のセットを生成し、トレーニング済み実モデル２４０を介して得られた知識を再作成する試みにおいて使用できるシミュレートされたデータセットまたはモンテカルロデータセットを生成する。プロキシデータ２６０を生成することにより、プライベートデータ２２２のトレーニングセットの匿名化の必要性を低減または排除できる。プロキシデータ２６０は、プライベートデータ２２２に格納されている実際の情報への参照を排除しながら、トレーニングデータの学習可能な顕著な特徴（すなわち知識）を保持するランダムに生成された合成データ、または場合によっては、決定論的に生成された合成データと見なすことができる。幾つかの実施形態で、プロキシサンプルが実際の患者データと著しく重複しないように、プロキシデータ２６０からのサンプルをプライベートデータ２２２内のサンプルと比較する。重複が著しいプロキシサンプルは、プライバシを確実に保護するた

めに破棄してもよい。プロキシサンプル除外フィルタは、プライベートデータ２２２内の患者データの名前空間（namespace）または属性空間に従って定義された基準に基づくことができる。例えば、プロキシサンプルは、１つ以上の実際のサンプルと共通する特徴（例えば、共通の郵便番号、共通の症状等）が多すぎる場合、除外できる。

【００７５】

幾つかの実施形態では、プロキシデータセットの生成中に、プロキシデータセットを決定論的に生成できるように、既知の「シード」を実装できる。したがって、モデルパラメータとシードをピアデバイスまたは非プライベートデバイスに送信して、プロキシデータの正確な複製を別の場所で生成できる。真の乱数のシードのソースは、例えば、ｒａｎｄｏｍ．ｏｒｇから入手できる（ＵＲＬ：www.random.orgを参照）。

【００７６】

幾つかの側面では、トレーニング済み機械学習モデルおよびプロキシデータ生成は、非可逆圧縮の一形態と見なすことができる。非可逆圧縮と同様に、元のデータをプロキシデータに変換すると、データの重要な特徴は維持されるが、個々の患者に関する詳細情報は、維持されない。カスタマイズされた圧縮の形式で一連のモデルパラメータを送達することによって、パラメータに基づいてデータを再作成できる。全てのプロキシデータセットをピアサーバに送信するのではなく、モデルパラメータ（事実上データセットの圧縮バージョンである、データ分布に基づく機械学習パラメータ）をシードと共に送信できる。ローカルマシンは、モデルパラメータとシードを受け取り、決定論的プロキシデータを再作成する。

【００７７】

したがって、プロキシデータの生成は、データサイエンス、人工知能、および分散コンピューティングの分野を明らかに向上させる。プロキシデータは、実際のデータの合成された等価物であるため、このデータは、大規模なデータセットと比較してよりコンパクトな形式で提供できる可能性があり、したがって、分散演算環境全体でデータセットを集約するために使用される人工知能プラットフォームの性能が向上する。例えば、何百万もの実際の個々のデータポイントに対する分布の形式でパラメータを提供することによって、人工知能プラットフォームが遥かに効率的に動作できるコンパクトなデータ表現が提供され、これによってシステムの全体的な機能が向上する。もちろん、ここに説明するように、プロキシデータは、これを用いなければ匿名化プロセス中に破棄される可能性がある知識を保存する。更に、プロキシデータを使用することによって、患者のプライバシおよびＨＩＰＰＡ規格への準拠を維持できる。

【００７８】

次に、モデリングエンジン２２６は、プロキシデータ２６０以外においてトレーニング済み実モデル２４０を作成するために使用された機械学習アルゴリズム２９５の同じ実装をトレーニングすることによって、プロキシデータ２６０からトレーニング済みプロキシモデル２７０を作成する。トレーニング済みプロキシモデル２７０は、プロキシモデルパラメータ２７５も含み、これは、実モデルパラメータ２４５とは僅かに異なる可能性が高い。幾つかの実施形態では、モデリングエンジン２２６は、トレーニング済みモデルのパラメータの少なくとも一部に基づいて、トレーニング済みプロキシモデル２７０がトレーニング済み実モデル２４０と十分に類似するまで、プロキシデータ２６０を繰り返し生成し、トレーニング済みプロキシモデル２７０を作成する。この手法は、２つのトレーニング済みモデルによって表されるように、プライベートデータ２２２から取得された知識を再現できる合成データを生成するために有利であると考えられる。

【００７９】

トレーニング済みプロキシモデル２７０とトレーニング済み実モデル２４０との間の類似性は、モデリングエンジン２２６が、プロキシモデルパラメータ２７５および実モデルパラメータ２４５の関数としてモデル類似性スコア２８０を算出することによって、様々な手法で測定できる。結果として得られるモデル類似性スコア２８０は、少なくとも類似性基準内で、２つのモデルがどれほど類似しているかを表すものである。類似性基準は、

プライベートデータ２２２の分析を要求する研究者によって定義でき、モデル指示２３０内で送達できる。幾つかの実施形態では、類似性スコア２８０は、後に閾値と比較できる単値（例えば、精度の差、二乗誤差の合計等）であってもよい。他の実施形態では、類似性スコア２８０は、多値であってもよい。例えば、多くのプロキシモデルが生成される場合、精度が正規分布に含まれると仮定して、類似性スコアは、幅と共に実モデルに対するプロキシモデルの精度の平均値を含むことができる。類似性スコア２８０が多値を含む実施形態では、類似性スコア２８０内の値は、類似性基準（すなわち複数の基準）と比較できる。類似性スコア２８０を測定するための技術は、図４を用いて後に更に説明する。

【００８０】

類似性スコア２８０が類似性基準を満たし、これによってトレーニング済みプロキシモデル２７０がトレーニング済み実モデル２４０と十分に類似していることを示す場合、モデリングエンジン２２６は、ここから得られた知識についての情報を送信できる。より具体的には、例えば、類似性基準が満たされると、モデリングエンジン２２６は、例えば、モデル指示２３０に従って、プロキシデータ２６０、プロキシモデルパラメータ２７５、類似性スコア２８０、または他の情報の１つまたは複数をネットワーク２１５を介して非プライベート演算デバイスに送信できる。上述のように、この手法によって、研究者は、プライバシまたはセキュリティを損なうことなく、プライベートデータ２２２についての知識を取得できる。

【００８１】

知識を受信する非プライベート演算デバイスは、その知識を他のプライベートデータサーバ２２４から得られた知識に集約できる。なお、非プライベート演算デバイス（例えば、図１の非プライベート演算デバイス１３０）は、エコシステム内の異なるプライベートデータサーバ、集中型機械学習ハブまたはサービス、グローバルモデリングエンジン、クラウドベースのサービス、またはデータを受信するように適切に構成された他のタイプの演算デバイスであってもよい。非プライベート演算デバイスとして動作する中央モデリングサービスの観点からは、中央モデリングサービスは、全てのプロキシデータセットを新しい集約されたトレーニングデータセットとして集約して、トレーニング済みグローバル集約モデルを作成できる。そして、集約されたモデルは、関心のある関係者、例えば、プライベートデータサーバ２２４に送り返され、患者の治療および転帰の分類器または予測器として使用される。更に、集約モデルは、トレーニング済み実モデル２４０の新しいバージョンのためのベースラインまたは基礎として使用できる。別の観点から言えば、モデル指示２３０は、グローバルトレーニング済みモデルを含むことができ、これをプライベートデータ２２２上で更にトレーニングして、トレーニング済み実モデル２４０を生成できる。グローバルトレーニング済みモデルをトレーニング済みプロキシモデル２７０の基礎としてもよい。

【００８２】

図３は、プライベートデータ分布の性質とプロキシデータの作成に関する更なる詳細を示している。プライベートデータ３２２は、トレーニング済み実モデルを作成するために使用されるトレーニングデータセットを表し、任意のまたは必要な前処理、例えば、プライベートデータ３２２の事実上の誤りの訂正が完了した後の入力データセットであると見なさる。プライベートデータ３２２内の各サンプルがデータの属性空間に応じて多くの値を含むことができる場合、プライベートデータ３２２は、多くの次元または属性を含むことができる。健康管理に関しては、プライベートデータ３２２は、以下に限定されるものではないが、ゲノムデータ、全ゲノム配列データ、全エキソソーム配列データ、プロテオームデータ、ネオエピトープデータ、ＲＮＡデータ、アレルギー情報、遭遇データ、治療データ、転帰データ、予約データ、オーダーデータ、請求コードデータ、診断コードデータ、結果データ、人口統計データ、投薬データ、バイタルサインデータ、支払人データ、薬物研究データ、薬物応答データ、経時的研究データ、バイオメトリックデータ、財務データ、所有権データ、電子医療記録データ、研究データ、人材データ、パフォーマンスデータ、分析結果データ、事象データ、またはその他のタイプのデータのうちの１つまたは

複数を含むことができる。したがって、プライベートデータ３２２内の単一のサンプルは、単一の患者、および公のまたはプライベートの患者の特定の属性または情報のセットを表すことができる。

【００８３】

プライベートデータ３２２内の全てのサンプルは、集約的に、トレーニングデータセット内の各関連次元に応じて、１つまたは複数のプライベートデータ分布３５０を形成する。例えば、プライベートデータ分布３５０は、年齢、体重、腫瘍配列における変異のタイプ、腫瘍ｖｓ対応正常ゲノム差分、または他の情報の分布を含むことができる。プライベートデータ分布３５０に関して「分布（distribution）」という表現を用いているが、分布には、多くの異なるタイプがある。例えば、性別の分布は、２つの数、すなわち、プライベートデータ３２２における女性の数および男性の数になる。この場合でも、プライベートデータ分布３５０は、ガウス分布、ポアソン分布、ベルヌーイ分布、ラーデマッヘル（Rademacher）分布、離散分布、二項分布、ゼータ分布、ガンマ分布、ベータ分布、ヒストグラム分布、または他のタイプの分布を含む明確に定義されたタイプの数学的または統計的分布を含むことができる。他の実施形態では、プライベートデータ分布３５０は、関連性のある次元の間で１つまたは複数の共分散行列を含むことができる。

【００８４】

他の実施形態では、データ分布を手動で構築してもよい（例えば、ヒストグラム、確率密度関数等）。幾つかの他の実施形態では、データ分布は、変化率および／または高次導関数（例えば、モーメント）に基づいてもよい。

【００８５】

明瞭さおよび説明の目的で、図３では、プライベートデータ分布３５０をＡおよびＢの２つの次元を有するグラフで表している。このグラフは、２つの次元の間に弱い相関があることを示している。これは、プライベートデータ分布３５０が、プライベートデータ３２２内の様々な属性または次元の間の他の相関を含むことがあることを実証するために提示しており、この相関は、プロキシデータ３６０を作成する際に保存することが好ましい。このような相関は、回帰、主成分分析、ピアソンの相関、ｋ－ｍｅａｎｓクラスタリング、またはトレーニングデータ内の次元間の関係を識別するために利用できるその他の手法によって発見することができる。

【００８６】

なお、プライベートデータ分布３５０は、プライベートデータ分布メタデータ３５０Ａによって示すように、追加の情報を含むことができる。メタデータ３５０Ａは、発見されたプライベートデータ分布３５０の性質に関する情報であり、カプセル化して他の演算デバイスに送信できる。メタデータの例は、分布の名称またはタイプ、分布を定義するパラメータ（例えば、平均値、最頻値、中央値、幅、シルエット係数、χ^2 フィット、ピアソン係数、モーメント等）、分布内のサンプル数、相関関係（例えば、主成分等）、またはプライベートデータ分布３５０を定義するために使用できる他の情報を含む。

【００８７】

プライベートデータ分布３５０は、プロキシデータ３６０を生成するために利用できる一種の確率分布と見なすことができる。プライベートデータ３２２に適合できる連続分布の場合、モデリングエンジンは、連続分布のメタデータ３５０Ａ（例えば、平均値、幅、モーメント等）を使用して、プロキシデータ３６０内の連続分布によってモデル化された次元に対して新しいサンプルの値をランダムに生成できる。不連続分布の場合、例えば、ヒストグラムは、関連する次元の値を生成するための離散確率密度関数として扱うことができる。例として、郵便番号（zip code）について検討する。患者データ３２２は、複数の郵便番号に亘る多数の患者サンプルポイントを含むことができる。関連する郵便番号についてのヒストグラムを生成し、このヒストグラムを正規化して郵便番号確率分布を生成できる。より具体的な例として、複数の郵便番号について、郵便番号のサブセットが特定のタイプの癌と関連付けられる場合、癌との相関を有する郵便番号からヒストグラムを構築できる。この郵便番号確率分布を反映して、合成患者データを構築できる。モデリング

エンジンは、正規化された郵便番号分布を使用して、プロキシデータ３６０の郵便番号値を生成する。

【００８８】

相関が発見されないまたは明確な相関がない場合、主成分分析（principle component analysis：ＰＣＡ）を利用してプライベートデータ３２２の次元を削減できる。次元を削減した後、新しいトレーニング済み実モデルを生成し、これを元のトレーニング済み実モデルと比較して、次元の削減後に知識が失われず、モデルの精度が維持されていることを確認する。データの次元を削減することにより、演算時間および伝送時間が短縮され、演算システムの性能が更に向上する。この比較は、図４を用いて説明する類似性スコアの手法を用いて行うことができる。モデリングエンジンは、次元を削減するためにデータにＰＣＡを適用でき、モデリングエンジンは、プライベートデータ３２２のための１つまたは複数の固有ベクトルまたは固有値を導出することもできる。「固有ベクトル（eigenvector）」は、トレーニングデータセットを表すために使用できる。したがって、プロキシデータ３６０は、プライベートデータ３２２、プライベートデータ分布３５０、実モデルパラメータ、またはプライベートデータ３２２に関連する他の情報から導出されるような固有ベクトルの組み合わせを含むと見なすことができる。例えば、プロキシデータ内の単一のサンプルは、固有ベクトルの線形の組み合わせ、場合によっては、重み付けされた組み合わせを含むことができる。このような組み合わせは、固有の患者（eigenpatient）、固有のプロファイル（eigenprofile）、固有の薬物（eigendrug）、固有の健康記録（eigen health record）、固有のゲノム（eigengenome）、固有のプロテオーム（eigenproteome）、固有のＲＮＡプロファイル（eigenRNA profile）、固有の経路（eigenpathway）、またはプライベートデータ３２２内のデータの性質に応じて他のタイプのベクトルを含むと考えることができる。

【００８９】

幾つかの実施形態では、固有値／固有ベクトルがペアで生じるように、各固有ベクトルは、ペアとなる固有値を有する。固有値は、データセット内の分散の尺度であり、固有ベクトルは、ｎ次元空間内のデータの方向を示す。与えられたデータセットに対して、固有値／固有ベクトルのペアの数は、データセットの次元数に等しくなる。ここに開示する技術に基づいて、このような情報の何れかおよび全てを利用できる。

【００９０】

プロキシデータ３６０には、複数の関心点（points of interest）が関連付けられる。プロキシデータ３６０は、必ずしもプライベートデータ３２２と同じ数のサンプルを有する必要はない。これに代えて、プロキシデータ３６０は、プロキシデータ３６０がトレーニング済み実モデルと同様のモデルを再現できる十分な数のサンプルを有すればよい。他の関心点として、図３に示すように、プロキシデータ３６０は、少なくとも許容範囲内または定義された範囲内で、異なるプロキシデータ分布３６２を有することができる。分布の差は、サンプルを作成する元となるプライベートデータ分布３５０と比較して、新しく生成されたサンプルにおけるランダム性のために僅かに異なる可能性がある。プロキシデータ分布３６２とプライベートデータ分布３５０との間の許容可能な差は、プロキシデータ３６０が所望の類似モデルを生成できることを保証するように調整できるハイパーパラメータと見なすことができる。更に、プロキシデータ３６０が、トレーニング済み実モデルと十分に類似するトレーニング済みプロキシモデルを生成する限り、分布間の差は、無視できるものとして許容できる。更に別の関心点は、所望の特徴、すなわち、許容可能なプロキシデータ分布３６２の特徴、許容可能な類似モデル、または他の因子を有するまで、プロキシデータ３６０を繰り返し生成できることである。例えば、モデリングエンジンは、遺伝的アルゴリズムを用い、適切な類似のトレーニング済みプロキシモデルが現れるまで、適合性関数として類似性スコアを使用してプロキシデータ３６０の値を変更でき、あるいは、実際のデータの共分散行列とプロキシデータ３６０の共分散行列との間の差を使用して、プロキシデータ３６０が実際のデータと同じまたは類似の形状を保持することを保証できる。トレーニング済み実モデルに適合するより優れたトレーニング済みプロキ

シモデルを実現するために、プロキシデータ分布３６２を調整または「変異（mutate）」させてもよい。

【００９１】

ここに説明するように、プロキシデータは、プライベートデータ分布に基づいており、プロキシデータがプライベートデータ分布を確実に反映することが重要である。例えば、５次元（各次元は、異なるタイプのデータ分布を表していてもよい）の場合、プロキシデータは、５タプルとして表すことができる。合成されたプロキシデータは、実際の患者と同様の特徴を有する偽の記録を有する「偽の」患者にマッピングされ、これを患者データと比較して、患者データの適切な表現であることを確認してもよい。

【００９２】

図４は、２つのトレーニング済みモデル、この例では、トレーニング済み実モデル４４０およびトレーニング済みプロキシモデル４７０の間の類似性スコア４９０を算出するための可能な手法を示している。トレーニング済み実モデル４４０は、上述のように、実世界の実際のプライベートデータ４２２についてトレーニングされている。トレーニング済みプロキシモデル４７０は、図３を用いて説明したように、プライベートデータ４２２のデータ分布の関数として構築された合成プロキシデータ４６０についてトレーニングされている。

【００９３】

トレーニング済みモデルのそれぞれは、トレーニング済みモデルを作成または再インスタンス化するために必要な特徴（例えば、パラメータ値、パラメータ数、層数、ノード数等）を定義する対応するモデルパラメータを含むと考えられる。モデルパラメータは、対応する機械学習アルゴリズムの基礎となる実装の性質に依存する。例えば、トレーニング済み実モデル４４０が２Ｄ　ＳＶＭを含む場合、実モデルパラメータは、ソフトマージンパラメータＣ、カーネル選択およびその値、閾値ε、切片、重み、または他のＳＶＭパラメータの値を含む可能性が高い。ニューラルネットワークの場合、実モデルパラメータは、層数、カーネル値、各層内のニューロン／ノードの数、学習率、運動量、エポック、入力の重み、またはニューラルネットワークを再インスタンス化することを可能にする他の値を含むことができる。

【００９４】

過剰適合を防ぐモデルパラメータも含めることができる。例えば、システムは、ローカルモデルの過剰トレーニングまたは過剰適合を防ぐために、定義されたモデル指示の一部として自動フィードバックを提供できる。例えば、多数のニューロンおよび複数の層を含むニューラルネット等のコンピューティング技術の進歩により、機械学習システムは、最適な適合を提供しない複雑なモデルを生成することがある。例えば、プライベートデータを最適に分類または特徴付けしない機械学習システムによって、線形または低次の適合ではなく、高次の適合（例えば、１２次多項式）が生成されることがある。過剰適合を防ぐために、機械学習モデルを生成するために使用されるモデル指示によって、ノード数、層数、アルゴリズムのタイプ等を制限してもよい。モデルを構築するためのデータ量が不十分であることも、過剰適合の原因になる。すなわち、モデルパラメータは、少数の（例えば、１０以下の）パラメータ値を含むこともあり、非常に多数の（例えば、１００万を超える）パラメータ値を含むこともある。ここで、パラメータが１００万個ある場合でも、このデータは、１００万個のパラメータを導出するために使用されたデータセットを送信するよりも遥かに小さいことが理解される。なお、トレーニング済み実モデルとトレーニング済みプロキシモデル４７０とが同じ機械学習アルゴリズムの同じ基礎となる実装上に構築されるため、プロキシモデルパラメータ４７５は、実モデルパラメータ４４５と全く同じ数のパラメータを含むはずである。但し、様々なパラメータの値は、２つの定性的グラフによって表されるように、異なる場合がある。実モデルパラメータ４４５およびプロキシモデルパラメータ４７５が正確に同じ数のパラメータを有するため、これらは、一対一で互いに比較できる。ここに示す例では、比較は、パラメータ毎の差分を示す差分パラメータ４８０によって表される。トレーニング済みプロキシモデル４７０がトレーニング

済み実モデル４４０と完全に同一である場合、差分パラメータ４８０は、全てゼロになる。ここで、トレーニング済みプロキシモデル４７０がプロキシデータ４６０に基づいて構築されているため、ゼロ以外の差が生じることも予想される。したがって、ここに示す例では、少なくとも、実モデルパラメータ（Ｐａ）４４５およびプロキシモデルパラメータ（Ｐｐ）４７５の値の関数として類似性スコア４９０を算出することによって、２つのトレーニング済みモデルを比較でき、類似性スコア４９０において、Ｎは、パラメータの数に対応し、ｉは、ｉ番目のパラメータに対応する。

【００９５】

類似性スコア４９０は、様々な手法によって、および対応するモデル指示に概要が示されている研究者の目的に応じて算出できる。幾つかの実施形態では、類似性スコア４９０は、モデルパラメータ間の差（例えば、パラメータ差分４８０）に基づいて算出できる。例えば、類似性スコア４９０は、差の和または差の二乗和、パラメータ間のメトリック距離、共分散の差、共分散行列内の要素の差等を含むことができる。幾つかの状況では、差の和より差の二乗和が好ましいと考えられる。パラメータが大きく異なる定義を有することがある場合、各差が等しく寄与するようにまたはこれらの重要性に応じて寄与するように、値を正規化または重み付けしてもよい。また、類似性スコア４９０は、パラメータ差分４８０以外に、トレーニング済みモデルの他の側面に基づいて求めてもよい。差分以外の他の値としては、１つ以上の検証セット、モデル精度向上、データ分布、または他の基準に基づくモデル精度の比較を含むことができる。

【００９６】

なお、類似性スコア４９０は、単値として示しているが、類似性スコア４９０は、上述の差または差以外の値のうちの２つ以上を含む多値であってもよい。例えば、このスコアは、差の合計と、パラメータ間の差の平均、可能であれば正規化された値との両方を含むことができる。

【００９７】

類似性スコア４９０が所望の方向に向かって確実に推移するようにするために、プロキシデータ４６０およびトレーニング済みプロキシモデル４７０の作成を複数回繰り返すことによって類似性スコア４９０を追跡してもよい。この意味で、類似性スコア４９０は、プロキシデータを生成する際の適合度を表していると言える。類似性スコア４９０が類似性送信要件を満たす場合、モデリングエンジンは、他の補助的情報（例えば、プロキシモデルパラメータ４７５、実モデルパラメータ４４５、パラメータ差分４８０、類似性スコア４９０等）と共に、プロキシデータ４６０をリモートの非プライベート演算デバイスに送信できる。幾つかの実施形態では、非プライベート演算デバイスは、多数の分散ピアからのプロキシデータ４６０をグローバルモデルに集約するように構成可能なグローバルモデリングエンジンとして動作する。ここに開示する手法によって、プライベートデータ４２２の各セットから得られる知識を保持しながら、プライベートデータ４２２を確実に安全に保つことができる。

【００９８】

類似性スコア４９０を確立するための更に別の手法として、様々な関連データセットを使用してトレーニング済みプロキシモデル４７０に対して交差検証（cross validation）を実行してもよい。幾つかの実施形態では、交差検証は、トレーニング済み実モデル４４０のためのトレーニングセットとして使用されるプライベートデータ４２２の異なる部分を使用して実行できる。２つのトレーニング済みモデルが十分に類似している場合、トレーニング済みプロキシモデル４７０は、実データを使用して、許容可能な予測結果を生成する。他の実施形態では、プロキシデータ４２２をトレーニングデータセットおよび検証データセットに分割し、これらを用いて交差畳み込み検証（cross-fold validation）を行ってもよい。ここで、トレーニングデータ４２２は、トレーニング済み実モデル４４０を生成するために使用され、プライベートデータ分布は、プロキシデータ４６０を生成するために使用される。トレーニング用プロキシデータは、トレーニング済みプロキシモデル４７０を生成するために使用され、このトレーニング済みプロキシモデルを検証するた

めに検証用プロキシデータが提供される。更に、トレーニング済みプロキシモデル４７０は、可能であればプロキシモデルパラメータ４７５と共に、エコシステム内の他のモデリングエンジン（例えば、図１の他のモデリングエンジン１２６、非プライベート演算デバイス１３０、グローバルモデルエンジン１３６等）に送信できる。これらの演算デバイスは、それぞれの、可能であればプライベートの同様のトレーニングデータセットについて、トレーニング済みプロキシモデル４７０を検証することを試みることができる。検証デバイスのそれぞれが検証作業を完了すると、モデル類似性スコア４９０の評価および導出のために、検証結果が元のモデリングエンジンに返される。図５は、分散型オンライン機械学習のコンピュータ実装方法５００を示している。方法５００は、多くのプライベートデータセットから集約されたトレーニング済みグローバルモデルの構築に関連する。トレーニング済みグローバルモデルは、予測作業に使用するために各エンティティに送り返すことができる。

【００９９】

動作５１０は、（例えば、プライベートデータサーバ１２４からまたは中央／グローバルサーバ１３０から）モデル指示を受信し、少なくとも１つの機械学習アルゴリズムの実装に従って、ローカルプライベートデータの少なくとも一部からトレーニング済み実モデル２４０を作成するためのモデリングエンジンとして動作するプライベートデータサーバを構成することによって開始される。モデル指示は、１つ以上のプロトコルによって、ネットワーク（例えば、無線ネットワーク、パケット交換ネットワーク、インターネット、イントラネット、仮想プライベートネットワーク、セルラネットワーク、アドホックネットワーク、ピアツーピアネットワーク等）を介して受信できる。幾つかの実施形態では、モデル指示は、完全な自己完結型パッケージを表す。例えば、モデル指示は、トレーニングデータセットを生成するためのクエリとして使用できる所望のプライベートデータ特徴の定義と共に、ターゲット機械学習アルゴリズムのコンパイル済み実装を含むことができる。モデリングエンジンは、パッケージを受け取ると、適切に構成されている場合、保護されたコンテナ内でトレーニングを実行できる。他の実施形態では、モデル指示は、機械学習アルゴリズムのローカルに格納された実装へのポインタを提供する。更に、モデル指示は、モデリングエンジンがそのローカルトレーニングタスクを完了することを可能にする追加の情報、例えば、類似性基準、類似性スコア定義、ローカルデータベースからプライベートデータを選択するためのクエリ変換命令、ベースラインとして事前にトレーニング済みのモデル、またはその他の情報を含むことができる。例えば、研究者が、特定の腫瘍変異、例えば、一塩基多型（single nucleotide polymorphism：ＳＮＰ）を有する患者が特定の薬に反応するかを判定することに関心がある場合、研究者は、変異と薬物に基づいてクエリ基準を構築し、クエリ基準をモデル指示内にカプセル化できる。

【０１００】

動作５２０は、モデリングエンジンが、機械学習アルゴリズムの実装をローカルプライベートデータでトレーニングすることによって、モデル指示に応じておよびローカルプライベートデータの少なくとも一部の関数としてトレーニング済み実モデルを作成することを含む。モデリングエンジンは、モデル指示内に提供されたプライベートデータ選択基準に基づいてトレーニングデータサンプルを構築できる。モデリングエンジンは、必要に応じてローカルデータベースのインデックス作成／検索システムに適合するように適切なフォーマットを設定した後、データ選択基準をローカルプライベートデータベースに送信する。これにより得られるセットは、ターゲット機械学習アルゴリズムのトレーニングセットになる。ターゲットの機械学習アルゴリズムの実装をトレーニングセットでトレーニングすることは、アルゴリズムの重みの調整、アルゴリズムへの入力の重みの調整、適合基準の最適化、交差畳み込み検証の実行、事前トレーニング済みモデルの更新、システムの過適合防止の制約、またはその他の動作を含むことができる。これにより得られるトレーニング済み実モデルは、トレーニング済み実モデルを再インスタンス化するために使用できる実モデルパラメータを含む。

【０１０１】

動作５３０は、ローカルプライベートデータトレーニングセットから１つまたは複数のプライベートデータ分布を生成することを含み、プライベートデータ分布は、トレーニング済み実モデルを作成するために使用されるトレーニングセットを集約的に表す。プライベートデータ分布の形式または性質は、データの性質（例えば、連続的、離散的）に応じて変化する。幾つかの場合、データ分布は、１次元を表し、例えば、ヒストグラム、度数プロット、時変値、または他の１次元表現として表される。他の場合、データ分布は、２次元以上の関連次元（例えば、２Ｄ、３Ｄ等）を表す。より具体的には、高次元データ分布は、クラスタ、相関、等値線、密度プロット、散布図、または他のタイプの高次分布を含むことができる。様々なデータ分布は、適切なビニングを用いた値のヒストグラムの作成、データプロットの作成、データへの曲線の適合、散布図の作成、主成分の計算、回帰の計算、またはその他の動作を含む様々な手法によって生成できる。なお、分布を生成することは、トレーニングデータセットを表す確率密度関数または確率質量関数を生成することと見なすことができる。

【０１０２】

動作５４０は、１つまたは複数のプライベートデータ分布に基づいてプロキシデータのセットを生成することを含む。モデリングエンジンは、プライベートデータ分布を確率分布として活用し、そこからプロキシデータを生成できる。モデリングエンジンは、確率分布に基づいて新しいデータをランダムに生成することによって、新しいプロキシデータサンプルを生成できる。モデリングエンジンは、各サンプルと、関連する確率分布の範囲内の位置とを比較することによって、サンプルが実データの性質に準拠することを確認できる。動作５４０は、プロキシデータが、集約的に、同じ分布空間内に適切な形状を生成することを確実にするために複数回実行しまたは繰り返してもよい。プロキシデータは、提供されたデータトレーニングセットと同じ数のサンプルを含むことができるが、プロキシデータのサンプル数は、これより多くても少なくてもよい。プロキシデータの各サンプルをトレーニングデータからのサンプルと比較して、プロキシサンプルが元の実際のサンプルと類似しすぎていないかを確かめることができる。プロキシサンプルが実際のサンプルと類似しているまたは同じである場合は、プライバシを維持するためにプロキシサンプルを破棄してもよい。プロキシデータセットの生成は、実データの分布に基づいて実行されるモンテカルロシミュレーションを使用して行うことができ、ここで、決定論的な手法でプロキシデータを生成するためにシードを利用してもよい。

【０１０３】

動作５５０は、モデリングエンジンが、同じタイプまたは実装の機械学習アルゴリズムをプロキシデータでトレーニングすることによって、プロキシデータからトレーニング済みプロキシモデルを作成することを続ける。特に注目すべき点として、２つのモデルを正確に比較できることを保証するために、トレーニング済みプロキシモデルは、トレーニング済み実モデルを作成するために使用された機械学習アルゴリズムの同じ実装から作成される。モデリングエンジンは、トレーニング済みプロキシモデルがモデル指示に従って、トレーニング済み実モデルと十分に類似した手法でトレーニングされることを保証する。トレーニング済みプロキシモデルは、完成すると、トレーニング済みプロキシモデルおよびプロキシデータを表すプロキシモデルパラメータを有することになる。トレーニング済みプロキシモデルとトレーニング済み実モデルが、典型的に全く同じ機械学習アルゴリズムの実装に基づいていても、これらの結果として得られるパラメータ値（例えば、重み、カーネル等）は、異なる可能性がある。

【０１０４】

動作５６０において、モデリングエンジンは、プロキシモデルパラメータと実モデルパラメータとの関数としてモデル類似性スコアを算出する。上述したように、各モデルが機械学習アルゴリズムの同じ実装から構築されていること、およびプロキシデータがプライベートデータと同様の特徴を有することを考慮して、パラメータをペア毎に比較できる。プロキシモデルパラメータおよび実モデルパラメータを使用することに加えて、モデリングエンジンは、類似性スコアを算出する際に利用可能な他の因子を使用することもできる

。他の因子の例としては、モデルの精度、交差畳み込み検証、精度向上、感度、特異性、ペア毎の比較の分布（例えば、平均値、ゼロ付近の分布等）がある。幾つかの実施形態において、実際のプライベートデータトレーニングセットを用いて、プロキシモデルを交差検証することができる。実際のプライベートデータトレーニングセットに対するトレーニング済みプロキシモデルからの予測の精度が十分に高い（例えば、１０％、５％、１％、またはこれより近い）場合、トレーニング済みプロキシモデルは、トレーニング済み実モデルに類似すると見なすことができる。更に、類似性スコアが類似性基準を満たさない（例えば、閾値を下回る）場合、モデリングエンジンは、動作５４０から動作５６０を繰り返すことができる。

【０１０５】

類似性スコアが類似性基準を満たすという条件により、モデリングエンジンは、動作５７０に進むことができる。動作５７０は、可能であれば他の情報と共に、ネットワークを介して少なくとも１つの非プライベート演算デバイスにプロキシデータのセットを送信することを含む。非プライベート演算デバイスは、プライベートサーバまたはピアハブ、あるいはこの両方からのプロキシデータを集約する集中型ハブであってもよい。プロキシデータは、マークアップ言語（例えば、ＸＭＬ、ＹＡＭＬ、ＪＳＯＮ等）、ｚｉｐアーカイブ、または他のフォーマットでシリアル化されたファイル（例えば、ＨＤＦ５）としてネットワークを介して送信できる。実モデルパラメータ、プロキシモデルパラメータ、データ分布、類似性スコア、またはその他の情報を含むプロキシデータ以外の追加情報を、リモート演算デバイス、グローバルモデリングエンジン、またはピアマシン等に送信してもよい。モデルパラメータを提供することにより、リモート演算デバイスは、トレーニング済みモデルを再インスタンス化し、プライベートデータサーバのモデリングエンジンによって実行された作業をローカルで検証できる。実際のプライベートデータは、送信されないため、プライバシが保護される。

【０１０６】

グローバルモデリングエンジンまたはピアプライベートデータマシンによって実行される動作５８０は、異なるプライベートデータサーバからの２つ以上のプロキシデータセットを集約することを含む。集約されたプロキシデータセット（グローバルプロキシセット）は、所与の機械学習タスクに基づいて組み合わされ、最初に要求されたモデル指示に従って生成される。プロキシデータの各セットは、異なるプライベートデータ分布から生成される可能性があるが、対応するプライベートデータトレーニングセットは、同じ選択基準に従って構築される。例えば、喫煙者が肺癌の治療にどの程度反応するかについての予測モデルを構築することを研究者が望んだとする。この研究では、それぞれが独自のプライベートデータを有する多くの個別の病院でモデルを構築することが求められる。各病院は、喫煙者である患者、治療を受けた患者、およびこれらに関連する既知の転帰等、同じデータ選択基準を受け取る。各病院のローカルプライベートデータサーバは、それぞれのモデリングエンジンを介して、実際のトレーニングデータを基礎として使用し、同じデータ選択基準に基づいて独自のプロキシデータを構築する。次に、グローバルモデリングエンジンは、個々のプロキシデータセットを集約してグローバルトレーニングデータセットを作成する。動作５９０は、グローバルモデリングエンジンが、集約されたプロキシデータのセットでグローバルモデルをトレーニングすることを含む。グローバルモデルは、各エンティティのプライベートデータから得られた知識を統合する。幾つかの実施形態では、グローバルモデリングエンジンは、実モデルパラメータのセットを蓄積し、これらを単一のトレーニング済みモデルに組み合わせることによってトレーニング済みグローバルモデルを作成できる。このような手法は、単純化された線形アルゴリズム、例えば、線形ＳＶＭに対して実行可能であると考えられる。一方、より複雑な実施形態では、例えば、プロキシデータセットを使用するニューラルネットワークが優れていると考えられ、これは、個々のパラメータを数学的に組み合わせること（例えば、加算、平均化等）によって失われる可能性があるプロキシデータセットにおける潜在的な知識が保持されるためである。

【０１０７】
他の実施形態では、グローバルモデリングエンジンは、トレーニング済みグローバルモデルを１つまたは複数のプライベートデータサーバに送り返す。次に、プライベートデータサーバは、グローバルトレーニング済みモデルを利用して、ローカルの臨床意思決定ワークフローを支援してローカルの予測調査を行うことができる。更に、プライベートデータサーバは、グローバルモデルを継続的なオンライン学習の基礎として使用することもできる。このように、グローバルモデルは、新しいプライベートデータが利用可能になると、継続的な機械学習の基礎となる。新しいデータが利用可能になると、方法５００を繰り返してグローバルモデリングエンジンを向上させることができる。

【０１０８】
機械学習システムは、複数の入力（例えば、プライベートデータ）を受け取ることができ、機械学習プロセスを通して、最も重要な入力のサブセットを識別できる。したがって、所与の病院は、他の病院と全く同じタイプのプライベートデータを収集しなくてもよい。すなわち、モデル指示は、病院や施設によって異なっていてもよい。しかしながら、ここに説明する機械学習システムを使用して、どのパラメータが最も予測的であるかを識別することによって、重要な予測パラメータを共通に有するデータセットを組み合わせることができる。他の実施形態では、モデル指示を修正し、例えば、主要な予測機能を含むように制限し、プロキシデータ、プロキシデータ分布、および他のタイプの学習情報を再生成するために使用できる。この再生成された情報は、後にグローバルモデルサーバに送信され、そこで集約される。

【０１０９】
他の実施形態では、第１の病院は、第２の病院とは異なる手法でデータを収集またはフィルタリングできる。したがって、データセットを組み合わせることができるようになる前に、必要とされるデータが別様に正規化されることがある。

【０１１０】
他の実施形態では、研究者は、特定のプライベートデータのセットに対して異なる分析を実行することを望む場合がある。例えば、モデル指示の第１のセットは、モデルを構築するためにガウス分布を使用するべきであると示すことがある。モデル指示の第２のセットは、モデルを構築するためにポアソン分布を使用するべきであると示すことがある。これらの結果を比較し、最も予測的なモデルを選択できる。結果を比較して、特定の機械学習モデルの再現性を評価することもできる。

【０１１１】
更に他の実施形態では、モデル指示の第１のセットを使用して特定のタイプの癌を研究し、例えば、乳癌分類子を作成できる。次に、モデル指示を修正し（例えば、更なる指示を追加し、乳癌に固有の指示を削除し、および前立腺癌に固有の指示を追加し）、異なる癌コホート、例えば、前立腺癌コホートでこのモデル指示を使用してもよい。このように、第１のタイプの癌についての第１の組のモデル指示は、幾つかの修正を加えて別のタイプの癌に外挿してもよいと考えられる。したがって、ここに開示する技術に基づいて、異なるタイプの癌とこれらの治療との間の新規な関係を検出できる。例えば、第１のタイプの癌と第２のタイプの癌との間には、相関関係が存在することがあり、この結果、第１のタイプの癌を治療することによって、第２のタイプの癌の治療の成功が予測される。

【０１１２】
図６は、方法５００の代替となる分散型オンライン機械学習のコンピュータ実装方法６００を示している。プライベートデータサーバ内のモデリングエンジンによって行われる動作６１０、６２０、６３０は、モデリングエンジンによって行われる動作５１０、５２０、５３０と同じである。方法６００は、動作６４０において方法５００とは実質的に異なっているが、初期的には、エンティティのプライベートデータサーバ内に展開されているモデリングエンジンのアクティビティに注目している点に変わりはない。方法６００は、リモートの非プライベート演算デバイスが、プライベートエンティティからのローカルプライベートデータを表すデータ分布からグローバルモデルを作成することを可能にする

。

【０１１３】

動作６４０は、モデリングエンジンがローカルプライベートデータ分布から１つまたは複数の顕著なプライベートデータ特徴を識別することを含む。顕著なプライベートデータ特徴とは、データ分布をモデル化するためまたは非プライベート演算デバイスのメモリ内で分布をインスタンス化するために必要なデータと見なすことができる。分布の性質に応じて、顕著なプライベートデータ特徴には、サンプル数、主成分、平均、最頻値、中央値、分布タイプ（例えば、ガウス、ポアソン、減衰等）、分布タイプパラメータ、ヒストグラムビニング、モーメント、相関、または他の特徴のうちの１つまたは複数を含めることができる。更に、好ましくは、顕著なプライベートデータ特徴は、実際のプライベートデータについてトレーニングされたトレーニング済み実モデルのパラメータを含むことができる。実モデルパラメータは、以下の手順で使用される。顕著なプライベートデータ特徴は、マークアップ言語（例えば、ＸＭＬ、ＹＡＭＬ、ＪＳＯＮ等）または他の任意の適切なフォーマットに基づいて送信用にパッケージ化できる。

【０１１４】

動作６５０は、モデリングエンジンがネットワークを介してリモート演算デバイスに顕著なプライベートデータ特徴を送信することに注目する。典型的な実施形態では、顕著なプライベートデータ特徴は、多くのプライベートエンティティからのこのような顕著な特徴を集約するグローバルモデリングエンジンに送信される。特徴の送信は、ＨＴＴＰ、ＨＴＴＰＳ、ＵＤＰ、ＴＣＰ、ＦＴＰ、ウェブサービス（例えば、ＲＥＳＴ、ＷＳＤＬ、ＳＯＡＰ等）、または他のプロトコルを含む１つまたは複数のネットワークプロトコルに基づいて行うことができる。

【０１１５】

動作６６０では、注目点がエンティティのプライベートデータサーバ内のモデリングエンジンから非プライベート演算デバイスのグローバルモデリングエンジン（例えば、図１のグローバルモデリングエンジン１３６）に移る。グローバルモデリングエンジンは、顕著なプライベートデータ特徴を受け取り、メモリ内でプライベートデータ分布を局所的に再インスタンス化する。動作５４０に関して上述したように、グローバルモデリングエンジンは、例えば、再インスタンス化されたプライベートデータ分布を確率分布として使用して、新しい合成サンプルデータを生成することによって、顕著なプライベートデータ特徴からプロキシデータを生成する。生成されたプロキシデータは、元の実際のデータと同数のサンプルを有する必要はない。プロキシデータは、単に、実際のプライベートデータから作成されたトレーニング済みモデルと同様のトレーニング済みモデルを作成するために十分な品質を有する十分な数のサンプルを有していればよい。

【０１１６】

幾つかの実施形態では、顕著なプライベートデータ特徴は、プロキシデータを含むことができ、このプロキシデータからデータ分布を再導出できる。しかしながら、方法６００の例では、各プライベートデータサーバは、自ら顕著な特徴を生成することが有利であると考えられる。この理由の１つは、グローバルなモデリングエンジンがプロキシデータセットに対する全ての作業を一元化されたシリアル形式で実行するのではなく、各プライベートデータサーバのモデリングエンジンが並列且つ分散形式で動作できるためである。したがって、システム全体のスループットが向上する。なお、プロキシデータが疎である場合、疎であるプロキシデータは、パッケージ化された顕著なプライベートデータ特徴よりもコンパクトに送信できるため、グローバルモデリングエンジンがプロキシデータを受け取ることが妥当である。プロキシデータを送信するか否かを決定するための条件または要件は、元のモデル指示にパッケージ化できる。

【０１１７】

図５の動作５５０と同様に、動作６７０において、グローバルモデリングエンジンは、トレーニング済み実モデルを作成するために使用されたものと同じタイプまたは実装の機械学習アルゴリズムをトレーニングすることによって、プロキシデータのセットからトレ

ーニング済みプロキシモデルを作成する。この場合、プロキシデータがトレーニングデータセットになる。トレーニング済みプロキシモデルのトレーニングが完了すると、トレーニング済みモデルを定義するプロキシモデルパラメータのセットが取得される。上述したように、プロキシモデルパラメータは、ターゲット演算デバイス（例えば、プライベートデータサーバ）のメモリ内のトレーニング済みプロキシモデルを再インスタンス化するために使用できる。

【０１１８】

図６の動作５６０と同様に、動作６８０は、グローバルモデリングエンジンが、プロキシモデルパラメータおよび実際のプロキシモデルパラメータの関数として、トレーニング済み実モデルに対するトレーニング済みプロキシモデルのモデル類似性スコアを算出することを含む。実際のプロキシモデルパラメータは、動作６４０に関して説明したように、顕著なプライベートデータ特徴と共に取得でき、あるいは、プロキシデータのモデリングエンジンに要求を送信して取得できる。モデル類似性スコアが類似性要件を満たすことができない場合、グローバルモデリングエンジンは、十分に類似したトレーニング済みプロキシモデルが生成されるまで、動作６６０から動作６８０を繰り返すことができる。

【０１１９】

動作６９０は、グローバルモデリングエンジンが、トレーニング済みプロキシモデルが類似性要件を満たすと判定したときに、プロキシデータを集約グローバルモデルに集約することを含む。プロキシデータは、他のプライベートデータサーバからの他のプロキシデータと集約されて集約モデルが作成される。なお、この手法は、時間の経過に伴って集約されたグローバルモデルが新しいプロキシデータによって継続的に更新されるオンライン学習法で実行できる。

【０１２０】

更に興味深い実施形態として、生成されたグローバルモデルをプライベートデータサーバ内のモデリングエンジンに送信して、予測目的に使用してもよい。更に、プライベートデータサーバがトレーニング済み実モデルを構築するための基礎としてグローバルモデルを活用することもできる。プライベートデータサーバは、十分な量のデータサンプルを欠いていたとしても、依然として知識の発見に貢献できる可能性があるので、この手法は、有利であると考えられる。

【０１２１】

ここに開示する分散型オンライン機械学習の手法は、トレーニング済みモデルを検証するための多数の技術を活用できる。１つの手法は、第１のプライベートデータサーバがトレーニング済み実モデルを他のプライベートデータサーバに送信することを含む。他のプライベートデータサーバは、自らのローカルデータでトレーニング済み実モデルを検証し、その結果を第１のプライベートデータサーバに送り返すことができる。更に、グローバルモデリングエンジンは、集約されたプロキシデータのグローバルコレクションを使用して、トレーニング済み実モデルに対して一回以上の交差畳み込み検証ステップを実行することもできる。逆も真になる。グローバルモデリングエンジンは、１つまたは複数のプライベートデータサーバにグローバルモデルを送信して、各プライベートデータサーバのローカルデータについてグローバルモデルを検証させることができる。適切な分析を確実に行うために、同じデータ選択要件に従って選択されたデータセットに対して様々なモデルを検証する。

【０１２２】

ここに開示する発明の主題の更に別の興味深い側面は、データが蓄積されるにつれて様々なトレーニング済みモデルを経時的に管理できることである。各モデルは、特定のデータ要件を有する包括的な分散型研究タスクのメンバと見なすことができる。したがって、各モデルは、様々なモデリングエンジンがタスクベースでこれらのモデルを管理することを可能にするタスク識別子（例えば、名称、目標、ＧＵＩＤ、ＵＵＩＤ等）と関連付けることができる。新規のモデルを作成することに加えて、モデリングエンジンは、トレーニングされた各モデルまたはプロキシデータを経時的に保持できる。新しいデータがプライ

ベートデータサーバから可視になると、モデリングエンジンは、可能であればタスク固有のリスナを介してデータを検出し、関連する新しいデータを実際のトレーニング済みデータに統合できる。更に、これに応じて関連するデータ分布を更新できる。幾つかの場合、新しいプロキシデータセットが生成され、この場合、以前に生成されたプロキシデータに追加できる新しいサンプルのみが生成される。したがって、本発明の主題は、モデリングエンジンが研究課題に関連するモデルを管理する時変モデル管理規則を確立することを含む。規則の例には、更新の報告、経時的なモデルパラメータの監視、既存のモデルまたは調査タスクの項目化、モデルまたはデータの変更に応じたアラートの生成、グローバルモデルサーバ（例えば、グローバルモデリングハブ、グローバルモデリングエンジン等）からの失われたモデルの回復、モデリングまたは研究タスクのログ化、モデルの保護、並びにこの他の管理機能等が含まれる。

【０１２３】

ここに開示するエコシステム／システムは、各演算デバイス（例えば、グローバルモデリングエンジン、プライベートデータサーバ等）が１つまたは複数のモデリングエンジンを有する、多くの演算デバイスに亘って分散したオンライン学習を提供する。モデリングエンジンは、多くのモデリングタスクを管理するように設定可能である。したがって、能動モデルの数は、数百、数千、更には、百万を超えるモデルになる可能性がある。したがって、本発明の主題は、分散システムにおける多数のモデルオブジェクトの管理装置または方法も含む。例えば、システムによる管理のために、各モデリングタスクに１つ以上の識別子または他のメタデータを割り当てることができる。より具体的には、識別子は、一意のモデル識別子、同じタスクに属するモデル間で共有されるタスク識別子、モデル所有者識別子、タイムスタンプ、バージョン番号、エンティティまたはプライベートデータサーバの識別子、ジオスタンプ、またはこの他のタイプのＩＤを含むことができる。更に、グローバルモデリングエンジンは、各プロジェクトのステータスをまとめて提示するダッシュボードを研究者に提示するように構成できる。ダッシュボードは、特定のモデルおよびこれらの現在の状態（例えば、ＮＵＬＬ、インスタンス化済み、トレーニング中、トレーニング済み、更新済み、削除済み等）にドリルダウンするように構成できる。

【０１２４】

ここに開示する技術は、多くの可能な用途を有する。本開示の主題は、変異を有する患者の治療および転帰に関するトレーニングモデルに主に注目しているが、他の可能な用途もある。主な用途の１つとして、結果として得られるローカルのトレーニング済み実モデルまたはグローバルモデルは、臨床試験の推奨システムの基礎になり得る。様々な患者および臨床試験中の薬物を含む薬物の治療および転帰データでトレーニングされた、多数のトレーニング済みモデル、実際のローカルモデル、またはグローバルモデルを想定する。新たな患者が様々な疾患（例えば、癌等）と診断されると、患者の医療施設に配置されたモデリングエンジンは、利用可能な関連するトレーニング済み実モデル、トレーニング済みプロキシモデル、またはトレーニング済みグローバルモデルに、患者のデータ（例えば、ＷＧＳ、ＷＥＳ、ゲノム差分オブジェクト、症状、人口統計等）を提出できる。トレーニング済みモデルは、そのモデルが元々トレーニングされている特定の治療に患者が反応するか否かの予測をもたらす。患者が現在試験中の治療に反応するとモデルが予測する場合、システムは、患者が候補となる可能性のある臨床試験のランク付けされたリストを提示でき、このリストは、例えば、予測の信頼度に従ってランク付けされている。可能性のある患者－試験の一致が検出されると、モデリングエンジンは、アラートまたは他の通知を生成でき、その通知は、１または複数の患者ケア関係者に送達される。更に、患者の治療に応じて、トレーニング済みモデルが追加のトレーニングを通して確実に更新されるようにするために、トレーニング済み実モデル、プロキシモデルおよびグローバルモデルにこれらのデータをフィードバックしてもよい。

【０１２５】

本発明の主題の更に別の興味深い側面は、異常値事象の発見の機会の提供である。新しい患者のデータがシステムに入力され、システム内のモデリングエンジンが患者について

可能な治療の結果を予測するシナリオを再考する。更に、この患者は、例えば、特定のゲノム変異体に基づけば、特定の治療に対してノンレスポンダーであると予測されると想定する。しかし、患者は、後にレスポンダーであることが判明したとする。１つ以上のモデリングエンジンが予測と実際の結果との間の有意差を検出すると、モデリングエンジンは、トレーニング済みモデルを所有または管理する研究者に通知を生成できる。このような異常値の検出により、幾つかの洞察が提供される。例えば、異常値は、１つ以上のトレーニング済みモデルの弱点を示している可能性がある。更に、異常値は、トレーニングデータセットに対する異常値において、どのような差分（例えば、他のゲノム差分等）によってデータが異常値となるのかを判定するために更に研究するべき真の異常値である可能性がある。自動異常値検出または発見により、更なる研究の可能性が提供される。

【０１２６】

ここに開示する技術は、健康管理以外に、例えば、コンピュータゲーム開発に関するＡＩ研究にも活用できる。このような場合、コンピュータゲームコンソール（例えば、ＰＳ４、Ｘ-Ｂｏｘ、ＰＣ等）は、上述したように、ゲーム固有のモデリングエンジンを用いて構成できる。個々のプレイヤがゲームをプレイすると、モデリングエンジンは、プレイヤと所与のシナリオとのインタラクションを観察し（すなわち、入力を収集し）、プレイヤの成功（すなわち結果）を検出して、ローカルのトレーニング済み実モデルのトレーニングデータを生成する。上記の技術を使用して、多くのプレイヤからプロキシデータを生成して、グローバルトレーニングモデルを作成できる。グローバルトレーニングモデルは、ゲームのＡＩの基礎となる。ゲームのＡＩは、グローバルのトレーニング済みモデルを使用して、新しいプレイヤの可能な次の動きを予測して、次に何が起こるかを予測できる。そして、ゲームは、戦術や戦略を適宜変更して、より挑戦的なゲームプレイを生成できる。健康管理およびゲーム以外にここに開示する技術を利用できる他の産業分野として、保険契約分析、消費者取引分析、通勤交通分析、または安全を維持することが要求される質の高いトレーニングデータを大量に有する他のタイプの分析が含まれる。

【０１２７】

ここに開示する発明の主題が有用な更に他の可能な分野として、プライベート画像コレクションからの学習がある。例えば、多くの個人の個々の家庭用コンピュータ上に、プライベート画像の複数の分散キャッシュがあると想定する。ここに開示する技術により、研究者またはデータ分析者は、特定の画像にアクセスする必要なく、プライベート画像コレクション内の情報を研究できる。このような特徴は、所有者の許可が得られていると仮定して、各個人のコンピュータにモデリングエンジンをインストールすることによって達成できる。モデリングエンジンは、オリジナル画像の形式のローカルトレーニングデータを、モデリングの指示に従って定義されている他のトレーニング情報（例えば、注釈、分類、シーンの説明、場所、時間、設定、カメラの向き等）と共に受信できる。モデリングエンジンは、オリジナルの画像とトレーニング情報からローカルのトレーニング済み実モデルを作成できる。プロキシデータは、例えば、トレーニング済み実モデルの固有ベクトルに基づいて、類似の画像を構成することによって生成できる。

【０１２８】

例えば、プライベート画像コレクションは、典型的には、地理的に異なる場所（例えば、異なる地域、都市、郵便番号、州等）における多数の異なる医院、医療用画像処理施設、または臨床／病理検査室に関連するまたはこれらの内にあるコンピュータまたはデータストレージ設備に存在する場合がある。このような場合、画像コレクションは、特定の患者並びにこれらのそれぞれの診断および治療履歴に関連する様々なスキャン（例えば、ＰＥＴ、ＳＰＥＣＴ、ＣＴ、ｆＭＲＩ等）を含む。あるいは、画像は、関連する患者情報に関連付けられた組織切片（典型的には、染料、フルオロフォア、または他の手法で光学的に検出可能な実体によって染色される）または免疫組織化学的に処置された切片を含むことができる。更に想定される画像は、同様に関連する患者情報に関連付けられた超音波画像（例えば、２Ｄ、３Ｄ、ドップラー）または映像、または血管造影画像または映像を含む。

【０１２９】
以上に説明したように、ここに提案する分散型学習システムは、多数の利点を提供することは明らかである。例えば、分散学習システムでパターンの大規模分析（例えば、顕微鏡における単一視野の分析に代わる全組織切片の画像取得）を行うことができ、これにより、人間ではできないような著しく大きなデータセットの処理が可能になる。更に、学習プロセスのために対応する多数の対応する画像または映像が利用可能であるため、分散学習システムは、臨床医によって直感的な手がかりとしてしか認識されないことが多いパラメータを特定できる。更に、このようにして学習された情報は、患者の身元および状態を損なうことなく、分散学習システムに情報送達可能に接続されている大規模な加入者ネットワークに亘って共有できる。

【０１３０】
これに代えて、人間の視点から類似の画像を生成するのではなく、モデリングエンジンは、例えば、遺伝的アルゴリズムを使用して、コンピュータが理解可能な特徴（例えば、記述子、キーポイント等）を有する合成画像を生成できる。記述子は、単値であっても多値であってもよく、類似性スコア（ヒストグラム記述子）を含んでいてもよい。説明のため、ＳＩＦＴ記述子（２０００年３月６日に出願され、Ｌｏｗに付与された米国特許第６，７１１，２９３号、発明の名称「Method and Apparatus for Identifying Scale Invariant Features in an Image and Use of Same for Locating an Object in an Image」参照）および画像所有者によって提供された分類情報に基づいてモデリングエンジンがトレーニングされると仮定する。プロキシ画像は、ランダムに重なり合う半透明の多角形を使用して生成できる。遺伝的アルゴリズムは、多角形のパラメータを変化させることができ、そして、元の画像のＳＩＦＴ記述子に対するプロキシ画像のＳＩＦＴ記述子を適合度の尺度として使用できる。プロキシ画像のＳＩＦＴ記述子が十分に類似している（例えば、分布、値、記述子の数等が類似している）場合、プロキシ画像が完成する。なお、この手法は、結果として得られるプロキシ画像が機械によって理解可能であるが人間には理解できないために有利であると考えられ、これにより、遠隔の非プライベート演算デバイスがプロキシ画像から学習を行うことができる。他の実施形態では、学習モデルのパラメータをシステムに提供でき、システムは、対応する記述子を生成できる。

【０１３１】
ここに提示する方法は、上述した特定の動作順序に限定されない。同様の多くの順序および変形が可能であることは、当業者にとって明らかである。

【０１３２】
本明細書では、システム１００全体、プライベートデータサーバ、ピアデータサーバ、およびグローバルモデリングエンジン等を含む、多くの異なる実施形態が想到される。上述の実施形態の少なくとも幾つかを包含する特許請求の範囲を表１に示す。
＜表１＞
１．ローカルプライベートデータにアクセスするように構成され、少なくとも１つのモデリングエンジンを含むプライベートデータサーバを使用してプロキシデータを生成するコンピュータ実装方法において、前記少なくとも１つのモデリングエンジンは、
前記プライベートデータから、機械学習アルゴリズムを使用してトレーニング済み実モデルを作成し、
前記ローカルプライベートデータの少なくとも一部から、ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成し、
前記複数のプライベートデータ分布に基づいて、プロキシデータのセットを生成し、
前記プロキシデータのセットから、前記機械学習アルゴリズムを使用して、トレーニング済みプロキシモデルを作成するように構成されている方法。
２．請求項１において、前記トレーニング済みプロキシモデルの作成に使用される機械学習アルゴリズムは、前記トレーニング済み実モデルの作成に使用されるものと同じ機械学習アルゴリズムである方法。
３．請求項１において、前記プライベートデータサーバは、グローバルサーバからモデ

ル指示を受信し、前記ローカルプライベートデータの少なくとも一部から前記トレーニング済み実モデルを作成する方法。
４．請求項３において、前記トレーニング済み実モデルは、前記モデル指示および前記ローカルプライベートデータの少なくとも一部に基づいて作成され、前記機械学習アルゴリズムは、前記ローカルプライベートデータでトレーニングされる方法。
５．請求項１において、前記トレーニング済みプロキシモデルは、プロキシモデルパラメータを生成し、前記トレーニング済み実モデルは、トレーニング済み実モデルパラメータを生成する方法。
６．請求項５において、前記プライベートデータサーバは、前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数としてモデル類似性スコアを算出するように構成される方法。
７．請求項６において、前記プライベートデータサーバは、前記モデル類似性スコアの関数として、ネットワークを介して、前記プロキシデータのセットを少なくとも１つの非プライベート演算デバイスに送信するように構成される方法。
８．請求項１において、前記ローカルプライベートデータは、患者固有のデータを含む方法。
９．請求項１において、前記ローカルプライベートデータは、ゲノムデータ、全ゲノム配列データ、全エキソソーム配列データ、プロテオームデータ、プロテオミクス経路データ、ｋ－ｍｅｒデータ、ネオエピトープデータ、ＲＮＡデータ、アレルギー情報、遭遇データ、治療データ、転帰データ、予約データ、注文データ、請求コードデータ、診断コードデータ、結果データ、治療反応データ、腫瘍反応データ、人口統計データ、投薬データ、バイタルサインデータ、支払者データ、薬物研究データ、薬物反応データ、経時的研究データ、バイオメトリックデータ、財務データ、所有権データ、電子カルテデータ、研究データ、人材データ、パフォーマンスデータ、分析結果データ、または事象データを含むデータの少なくとも１つを含む方法。
１０．請求項１において、前記モデリングエンジンは、前記トレーニング済み実モデルを新しいローカルプライベートデータで更新するように更に構成される方法。
１１．請求項３において、前記モデル指示は、前記プライベートデータサーバの外部で作成されたベースラインモデルから前記トレーニング済み実モデルを作成するための指示を含む方法。
１２．請求項１１において、前記ベースラインモデルは、グローバルトレーニングモデルを含む方法。
１３．請求項６において、前記類似性スコアは、前記トレーニング済みプロキシモデルの交差検証に基づいて判定される方法。
１４．請求項１３において、前記交差検証は、
（１）前記プロキシデータの一部に対する内部交差検証、
（２）ローカルプライベートデータの一部の内部交差検証、または
（３）複数のプライベートデータサーバのうちの異なる１つによるそのローカルプライベートデータでの外部交差検証のうちの１つ以上を含む方法。
１５．請求項６において、前記類似性スコアは、
（１）前記プロキシモデルの正確度と前記トレーニング済み実モデルの正確度との差、または
（２）前記トレーニング済み実モデルパラメータおよび前記プロキシモデルパラメータを使用して算出されたメトリック距離のいずれかを含む方法。
１６．請求項７において、前記プロキシデータは、前記モデル類似性スコアの関数が少なくとも１つの送信基準を満たすときに送信される方法。
１７．請求項１６において、前記少なくとも１つの送信基準は、前記類似性スコアに関して、閾値条件、多値条件、値の変化条件、傾向条件、人的命令条件、外部要求条件、および時間条件の少なくとも１つを含む方法。
１８．請求項１において、前記プライベートデータを格納するローカルストレージシス

テムは、ローカルデータベース、ＢＡＭサーバ、ＳＡＭサーバ、ＧＡＲサーバ、ＢＡＭＢＡＭサーバ、および臨床オペレーティングシステムサーバのうち少なくとも１つを含む方法。

１９．請求項１において、前記複数のプライベートデータ分布の分布は、ガウス分布、ポアソン分布、ベルヌーイ分布、ラデマッハ分布、離散分布、二項分布、ゼータ分布、ガンマ分布、ベータ分布、およびヒストグラム分布の少なくとも１つに従う方法。

２０．請求項１において、前記プライベートデータ分布は、前記トレーニング済み実モデルパラメータと前記ローカルプライベートデータから導出された固有値に基づく方法。

２１．請求項１において、前記プロキシデータのセットは、前記トレーニング済み実モデルパラメータから導出された固有ベクトルと前記ローカルプライベートデータとの組み合わせを含む方法。

２２．請求項２１において、前記プロキシデータは、前記固有ベクトルの線形結合を含む方法。

２３．請求項２２において、前記固有ベクトルは、固有の患者、固有のプロファイル、固有の薬剤、固有の健康記録、固有のゲノム、固有のプロテオーム、固有のＲＮＡプロファイル、および固有の経路の少なくとも１つを含む方法。

２４．請求項１において、前記トレーニング済み実モデルは、分類アルゴリズム、ニューラルネットワークアルゴリズム、回帰アルゴリズム、決定木アルゴリズム、クラスタリングアルゴリズム、遺伝的アルゴリズム、教師あり学習アルゴリズム、半教師あり学習アルゴリズム、教師なし学習アルゴリズム、または深層学習アルゴリズムを含む機械学習アルゴリズムの少なくとも１つの実装に基づく方法。

２５．請求項１において、前記トレーニング済み実モデルは、サポートベクターマシン、最近傍アルゴリズム、ランダムフォレスト、リッジ回帰、Ｌａｓｓｏアルゴリズム、ｋ－ｍｅａｎｓクラスタリングアルゴリズム、スペクトルクラスタリングアルゴリズム、平均シフトクラスタリングアルゴリズム、非負行列因数分解アルゴリズム、エラスティックネットアルゴリズム、ベイズ分類アルゴリズム、ＲＡＮＳＡＣアルゴリズム、および直交マッチング追跡アルゴリズムを含む機械学習アルゴリズムの少なくとも１つの実装に基づく方法。

２６．複数のプライベートデータサーバと、少なくとも１つのグローバルモデリングエンジンを含むグローバルモデルサーバとを備える分散型機械学習システムにおいてプロキシデータを生成するコンピュータ実装方法であって、前記少なくとも１つのグローバルモデリングエンジンは、

前記クエリに基づいてモデル指示を生成し、

前記モデル指示を複数のプライベートデータサーバに送信し、

前記複数のプライベートデータサーバの各サーバからプロキシデータのセットを受信し、

前記プロキシデータのセットをグローバルプロキシデータに集約し、

前記グローバルプロキシデータを使用してグローバル集約モデルをトレーニングするように構成されている方法。

２７．請求項２６において、更に、

前記クエリに基づいて第１のモデル指示のセットを生成し、前記第１のモデル指示のセットを第１のプライベートデータサーバに送信することと、

前記クエリに基づいて、前記第１のモデル指示のセットとは異なる第２のモデル指示のセットを生成し、前記第２のモデル指示のセットを第２のデータサーバに送信することと、

第１のプライベートデータサーバから第１のプロキシデータのセットを受信し、第２のプライベートデータサーバから第２のプロキシデータセットを受信することと、

前記第１のプロキシデータのセットと前記第２のプロキシデータセットをグローバルプロキシデータに集約することとを含む方法。

２８．請求項２６において、更に、

前記プライベートデータサーバから、前記プライベートデータサーバに格納されているプライベートデータのタイプを示すメタデータを受信することと、
前記メタデータに基づいて前記モデル指示を生成することとを含む方法。
２９．請求項２６において、更に、前記グローバル集約モデルに基づいて、更新されたモデル指示を前記複数のプライベートデータサーバに提供することを含む方法。
３０．請求項２６において、前記グローバルトレーニングモデルは、少なくとも部分的に、前記複数のプライベートデータサーバのうちの少なくとも２つからのプロキシデータのセットでトレーニングされる方法。
３１．ローカルプライベートデータにアクセスするように構成され、少なくとも１つのモデリングエンジンを含むプライベートデータサーバを使用してプロキシデータを生成するコンピュータ実装方法において、前記少なくとも１つのモデリングエンジンは、
ピアプライベートデータサーバから、前記ピアプライベートデータサーバに格納されているピアプライベートデータに基づくピアプロキシデータのセットを受信し、
前記プライベートデータサーバに格納されているローカルプライベートデータから、機械学習アルゴリズムを使用してトレーニング済み実モデルを作成し、
前記ローカルプライベートデータの少なくとも一部から複数のプライベートデータ分布を生成し、
前記複数のプライベートデータ分布に基づいて、プロキシデータのセットを生成し、
前記ピアプロキシデータのセットを前記プロキシデータのセットと組み合わせて、集約プロキシデータのセットを形成し、
前記集約プロキシデータのセットから、前記プライベートデータサーバ上の機械学習アルゴリズムを使用して、トレーニング済みプロキシモデルを作成するように構成されている方法。
３２．請求項３１において、前記プライベートデータ分布は、前記ローカルプライベートデータを集約的に表す方法。
３３．請求項３１において、前記集約プロキシデータをグローバルモデリングエンジンに送信することを含む方法。
３４．請求項３１において、前記ピアプライベートデータサーバは、前記プライベートデータサーバ上の前記ローカルプライベートデータにアクセスする権限を有さない方法。
３５．分散型機械学習のコンピュータ実装方法であって、
プライベートデータサーバによって、前記プライベートデータサーバにローカルなローカルプライベートデータの少なくとも一部から、機械学習アルゴリズムの実装に従って、トレーニング済み実モデルを作成するためのモデル指示を受信することと、
機械学習エンジンによって、前記機械学習アルゴリズムの実装を前記ローカルプライベートデータでトレーニングすることによって、前記モデル指示に従って、および前記ローカルプライベートデータの少なくとも一部の関数として、トレーニング済み実モデルパラメータを含む前記トレーニング済み実モデルを作成することと、
前記機械学習エンジンによって、前記ローカルプライベートデータから、前記トレーニング済み実モデルを作成するために使用される前記ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成することと、
前記機械学習エンジンによって、前記プライベートデータ分布から、前記複数のプロキシデータ分布の複製を可能にする顕著なプライベートデータ特徴を識別することと、
前記機械学習エンジンによって、ネットワークを介して、前記顕著なプライベートデータ特徴を非プライベート演算デバイスに送信することとを含む方法。
３６．請求項３５において、前記顕著なプライベートデータ特徴は、プロキシデータのセットを含む方法。
３７．請求項３５において、前記複数のプライベートデータ分布および前記顕著なプライベートデータ特徴の少なくとも１つに従ってプロキシデータのセットを生成することを更に含む方法。
３８．請求項３７において、前記機械学習アルゴリズムの実装のタイプを前記プロキシ

データのセットでトレーニングすることによって、前記プロキシデータのセットから、プロキシモデルパラメータを含むトレーニング済みプロキシモデルを作成することを更に含む方法。

３９．請求項３８記載において、前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数として、前記トレーニング済みプロキシモデルのモデル類似性スコアを算出することを更に含む方法。

４０．請求項３９において、前記モデル類似性スコアに基づいて、前記プロキシデータのセットを集約グローバルモデルに集約することを更に含む方法。

４１．分散型機械学習システムにおいて、

ローカルプライベートデータを記憶する記憶デバイスと、

複数のプライベートデータサーバとを備え、前記複数のプライベートデータサーバは、ネットワークを介して通信可能に接続され、各プライベートデータサーバは、前記ローカルプライベートデータにアクセスするように構成され、各プライベートデータサーバは、１つまたは複数のプロセッサと、少なくとも１つのモデリングエンジンとを備え、前記少なくとも１つのモデリングエンジンは、

前記プライベートデータから、機械学習アルゴリズムを使用してトレーニング済み実モデルを作成し、

前記ローカルプライベートデータの少なくとも一部から、前記ローカルプライベートデータを集約的に表す複数のプライベートデータ分布を生成し、

前記複数のプライベートデータ分布に基づいて、プロキシデータのセットを生成し、

前記プロキシデータのセットから、前記機械学習アルゴリズムを使用して、トレーニング済みプロキシモデルを作成するように構成されているシステム。

４２．請求項４１において、前記トレーニング済みプロキシモデルの作成に使用される機械学習アルゴリズムは、前記トレーニング済み実モデルの作成に使用されるものと同じ機械学習アルゴリズムであるシステム。

４３．請求項４１において、前記プライベートデータサーバは、グローバルサーバからネットワークインタフェースを介してモデル指示を受信し、前記ローカルプライベートデータの少なくとも一部から前記トレーニング済み実モデルを作成するシステム。

４４．請求項４３において、前記トレーニング済み実モデルは、前記モデル指示に基づいておよび前記ローカルプライベートデータの少なくとも一部に基づいて作成され、前記機械学習アルゴリズムは、前記ローカルプライベートデータでトレーニングされるシステム。

４５．請求項４１において、前記トレーニング済みプロキシモデルは、プロキシモデルパラメータを生成し、前記トレーニング済み実モデルは、トレーニング済み実モデルパラメータを生成するシステム。

４６．請求項４５において、前記プライベートデータサーバは、前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数としてモデル類似性スコアを算出するように構成されるシステム。

４７．請求項４６において、前記プライベートデータサーバは、ネットワークインタフェースにより、前記モデル類似性スコアの関数として、前記プロキシデータのセットをネットワークを介して少なくとも１つの非プライベート演算デバイスに送信するように構成されるシステム。

４８．請求項４１において、前記モデリングエンジンは、前記トレーニング済み実モデルを新しいローカルプライベートデータで更新するように更に構成されるシステム。

４９．請求項４１において、前記モデル指示は、前記プライベートデータサーバの外部で作成されたベースラインモデルから前記トレーニング済み実モデルを作成するための指示を含むシステム。

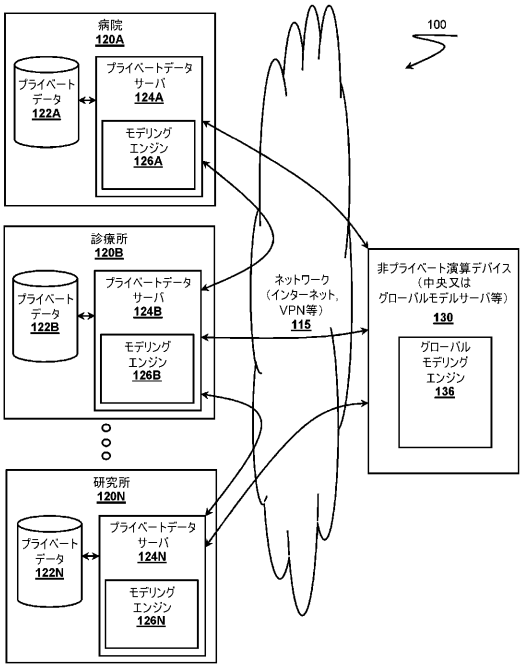
５０．請求項４９において、前記ベースラインモデルは、グローバルトレーニングモデルを含むシステム。

５１．請求項４７において、前記プロキシデータは、モデル類似性スコアが少なくとも

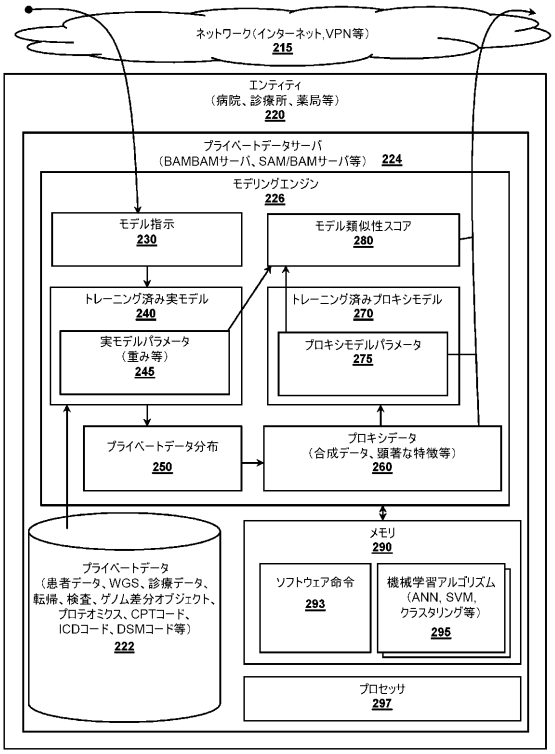
１つの送信基準を満たすときに送信されるシステム。
５２．請求項５１において、前記少なくとも１つの送信基準は、類似性スコアに関して、閾値条件、多値条件、値の変化条件、傾向条件、人的命令条件、外部要求条件、および時間条件の少なくとも１つを含むシステム。
５３．請求項４１において、前記ローカルプライベートデータを格納するローカルストレージシステムは、ローカルデータベース、ＢＡＭサーバ、ＳＡＭサーバ、ＧＡＲサーバ、ＢＡＭＢＡＭサーバ、および臨床オペレーティングシステムサーバのうち少なくとも１つを含むシステム。
５４．請求項４１において、前記トレーニング済み実モデルは、分類アルゴリズム、ニューラルネットワークアルゴリズム、回帰アルゴリズム、決定木アルゴリズム、クラスタリングアルゴリズム、遺伝的アルゴリズム、教師あり学習アルゴリズム、半教師あり学習アルゴリズム、教師なし学習アルゴリズム、深層学習アルゴリズム、サポートベクターマシン、最近傍アルゴリズム、ランダムフォレスト、リッジ回帰、Ｌａｓｓｏアルゴリズム、ｋ－ｍｅａｎｓクラスタリングアルゴリズム、スペクトルクラスタリングアルゴリズム、平均シフトクラスタリングアルゴリズム、非負行列因数分解アルゴリズム、エラスティックネットアルゴリズム、ベイズ分類アルゴリズム、ＲＡＮＳＡＣアルゴリズム、および直交マッチング追跡アルゴリズムを含む機械学習アルゴリズムの少なくとも１つの実装に基づくシステム。
５５．ネットワークを介して通信可能に接続された複数のプライベートデータサーバを含む分散型機械学習システム内のグローバルモデリングサーバであって、前記グローバルモデリングサーバは、１つまたは複数のプロセッサと、少なくとも１つのグローバルモデリングエンジンとを含み、前記少なくとも１つのグローバルモデリングエンジンは、
クエリに基づいてモデル指示を生成し、
ネットワークインタフェースを介して、前記モデル指示を複数のプライベートデータサーバに送信し、
前記ネットワークインタフェースを介して、前記複数のプライベートデータサーバからプロキシデータのセットを受信し、
前記プロキシデータのセットをグローバルプロキシデータに集約し、
前記グローバルプロキシデータを使用して前記グローバル集約モデルをトレーニングするように構成されているサーバ。
５６．請求項５５において、前記グローバルモデリングエンジンは、更に、
前記クエリに基づいて第１のモデル指示のセットを生成し、前記ネットワークインタフェースを介して、前記第１のモデル指示のセットを第１のプライベートデータサーバに送信し、
前記クエリに基づいて、前記第１のモデル指示セットとは異なる第２のモデル指示のセットを生成し、前記ネットワークインタフェースを介して、前記第２のモデル指示セットを第２のデータサーバに送信し、
前記第１のプライベートデータサーバから第１のプロキシデータのセットを受信し、前記ネットワークインタフェースを介して前記第２のプライベートデータサーバから第２のプロキシデータセットを受信し、
前記第１のプロキシデータのセットと前記第２のプロキシデータのセットをグローバルプロキシデータに集約するように構成されているサーバ。
５７．請求項５５において、前記グローバルモデリングエンジンは、更に、
前記ネットワークインタフェースを介して、プライベートデータサーバから、前記プライベートデータサーバに格納されているプライベートデータのタイプを示すメタデータを受信し、
前記メタデータに基づいて前記モデル指示を生成するように構成されているサーバ。
５８．請求項５５において、前記グローバルモデリングエンジンは、更に、
前記グローバル集約モデルに基づいて、前記ネットワークインタフェースを介して、更新されたモデル指示を前記複数のプライベートデータサーバに提供するように構成されて

いるサーバ。
５９．請求項５５において、前記グローバルトレーニングモデルは、少なくとも部分的に、前記複数のプライベートデータサーバの少なくとも２つからのプロキシデータのセットでトレーニングされるサーバ。
６０．ネットワークを介して通信可能に接続された複数のプライベートデータサーバとグローバルモデルサーバとを備える分散型機械学習システム内のプライベートデータサーバであって、前記プライベートデータサーバは、１つ以上のプロセッサと、少なくとも１つのモデリングエンジンとを含み、前記少なくとも１つのモデリングエンジンは、
ネットワークインタフェースを介して、ピアプライベートデータサーバから、前記ピアプライベートデータサーバに格納されているピアプライベートデータに基づくピアプロキシデータのセットを受信し、
前記プライベートデータサーバに格納されているローカルプライベートデータから、機械学習アルゴリズムを使用してトレーニング済み実モデルを作成し、
前記ローカルプライベートデータの少なくとも一部から複数のプライベートデータ分布を生成し、
前記複数のプライベートデータ分布に基づいて、プロキシデータのセットを生成し、
前記ピアプロキシデータのセットを前記プロキシデータのセットと組み合わせて、集約プロキシデータのセットを形成し、
前記集約プロキシデータのセットから、前記プライベートデータサーバ上の機械学習アルゴリズムを使用して、トレーニング済みプロキシモデルを作成するように構成されているプライベートデータサーバ。
６１．請求項６０において、前記プライベートデータ分布は、前記ローカルプライベートデータを集約的に表すプライベートデータサーバ。
６２．請求項６０において、前記プライベートデータサーバは、更に、ネットワークインタフェースを介して、前記集約プロキシデータをグローバルモデリングエンジンに送信するように構成されているプライベートデータサーバ。
６３．請求項６０において、前記ピアプライベートデータサーバは、前記プライベートデータサーバ上の前記ローカルプライベートデータにアクセスする権限を有さないプライベートデータサーバ。
【０１３３】
ここに説明した発明概念から逸脱することなく、上述したもの以外に更に多くの変形が可能であることは、当業者にとって明らかである。したがって、本発明の主題は、特許請求の範囲の趣旨以外によっては制限されない。更に、明細書と特許請求の範囲の両方を解釈する際に、全ての用語は、文脈と矛盾しない最も広い意味で解釈される。特に、「備える」、「有する」および「含む」といった用語は、要素、構成要素、またはステップを非排他的に指示するものとして解釈され、指示されている要素、構成要素、またはステップは、明示的に指示されていない他の要素、構成要素、またはステップと共に存在し、利用され、またはこれらと組み合わされてもよい。明細書または特許請求の範囲において、Ａ、Ｂ、Ｃ、…およびＮからなるグループから選択される少なくとも１つについて言及する場合、このテキストは、Ａ＋ＮまたはＢ＋Ｎ等ではなく、グループの１つの要素のみが要求されると解釈される。

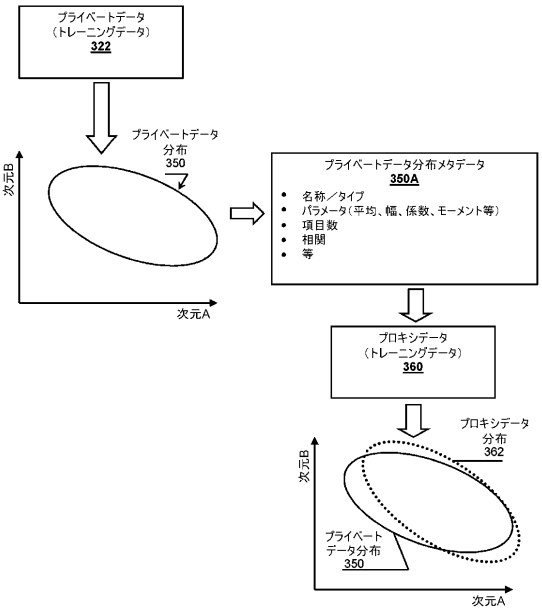
【図１】


病院
120A
プライベートデータ
122A
プライベートデータサーバ
124A
モデリングエンジン
126A
100
診療所
120B
プライベートデータ
122B
プライベートデータサーバ
124B
モデリングエンジン
126B
ネットワーク（インターネット，VPN等）
115
非プライベート演算デバイス（中央又はグローバルモデルサーバ等）
130
グローバルモデリングエンジン
136
研究所
120N
プライベートデータ
122N
プライベートデータサーバ
124N
モデリングエンジン
126N

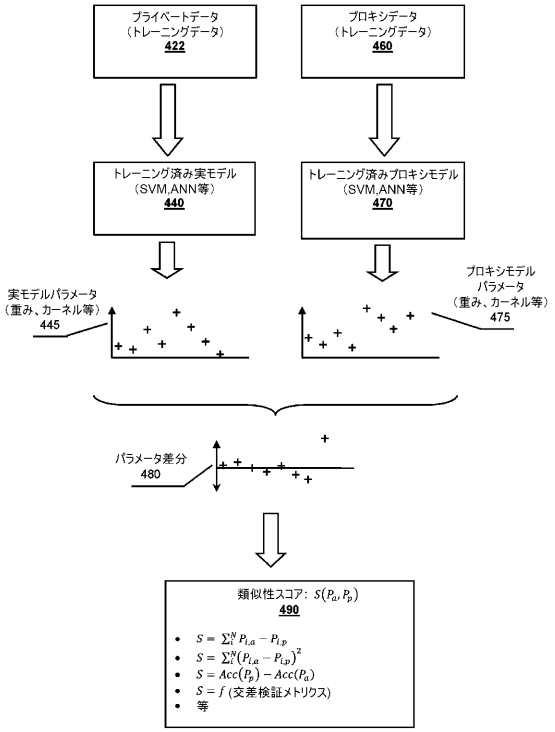
【図２】


ネットワーク（インターネット,VPN等）
215
エンティティ（病院、診療所、薬局等）
220
プライベートデータサーバ（BAMBAMサーバ、SAM/BAMサーバ等）
224
モデリングエンジン
226
モデル指示
230
モデル類似性スコア
280
トレーニング済み実モデル
240
実モデルパラメータ（重み等）
245
トレーニング済みプロキシモデル
270
プロキシモデルパラメータ
275
プライベートデータ分布
250
プロキシデータ（合成データ、顕著な特徴等）
260
プライベートデータ（患者データ、WGS、診療データ、転帰、検査、ゲノム差分オブジェクト、プロテオミクス、CPTコード、ICDコード、DSMコード等）
222
メモリ
290
ソフトウェア命令
293
機械学習アルゴリズム（ANN, SVM, クラスタリング等）
295
プロセッサ
297

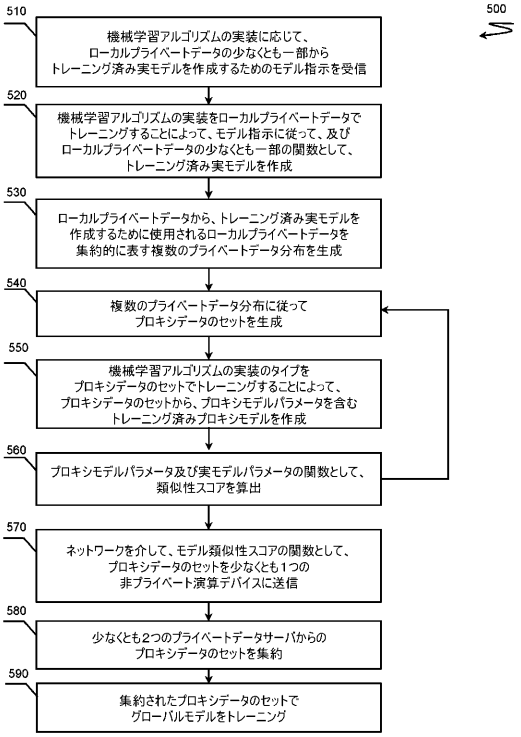
【図３】


プライベートデータ（トレーニングデータ）
322
プライベートデータ分布
350
次元B
次元A
プライベートデータ分布メタデータ
350A
名称／タイプ
パラメータ（平均、幅、係数、モーメント等）
項目数
相関
等
プロキシデータ（トレーニングデータ）
360
プロキシデータ分布
362
プライベートデータ分布
350

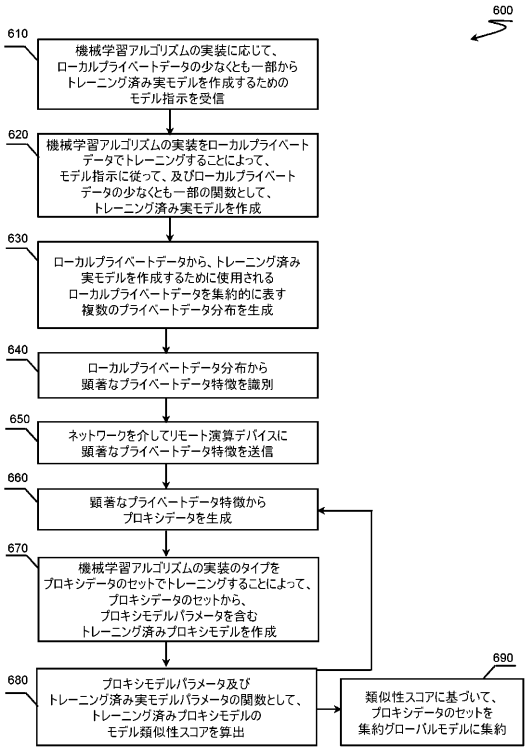
【図４】


プライベートデータ（トレーニングデータ）
422
プロキシデータ（トレーニングデータ）
460
トレーニング済み実モデル（SVM,ANN等）
440
トレーニング済みプロキシモデル（SVM,ANN等）
470
実モデルパラメータ（重み、カーネル等）
445
プロキシモデルパラメータ（重み、カーネル等）
475
パラメータ差分
480
類似性スコア：$S(P_a, P_p)$
490
$S = \sum_i^N P_{i,a} - P_{i,p}$
$S = \sum_i^N (P_{i,a} - P_{i,p})^2$
$S = Acc(P_p) - Acc(P_a)$
$S = f$（交差検証メトリクス）
等

【図５】



【図６】



【手続補正書】
【提出日】平成31年2月20日(2019.2.20)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
分散型機械学習システムであって、
それぞれがローカルプライベートデータへのアクセスを有しおよび少なくとも１つのモデリングエンジンを有する複数のプライベートデータサーバを備え、前記複数のプライベートデータサーバは、ネットワークを介して、少なくとも１つの非プライベート演算デバイスに通信可能に接続されており、
各前記プライベートデータサーバは、非一時的コンピュータ可読メモリに格納された少なくとも１つのプロセッサソフトウェア命令による実行に応じて、その少なくとも１つのモデリングエンジンに、
前記ローカルプライベートデータの少なくとも一部から、機械学習アルゴリズムの実装に従って、トレーニング済み実モデルを作成するためのモデル指示を受信するステップ、
前記機械学習アルゴリズムの実装を前記ローカルプライベートデータでトレーニングすることによって、前記モデル指示に従っておよび前記ローカルプライベートデータの少なくとも一部の関数として、トレーニング済み実モデルパラメータを含む前記トレーニング済み実モデルを作成するステップ、
前記ローカルプライベートデータから、前記トレーニング済み実モデルを作成するために使用される前記ローカルプライベートデータを集約的に表す複数のプライベートデータ

分布を生成するステップ、
前記複数のプライベートデータ分布に従ってプロキシデータのセットを生成するステップ、
前記機械学習モデルのタイプを前記プロキシデータのセットでトレーニングすることによって、前記プロキシデータのセットから、プロキシモデルパラメータを含むトレーニング済みプロキシモデルを作成するステップ、
前記プロキシモデルパラメータおよび前記トレーニング済み実モデルパラメータの関数としてモデル類似性スコアを算出するステップ、
前記ネットワークを介して、前記モデル類似性スコアの関数として、前記プロキシデータのセットを少なくとも１つの非プライベート演算デバイスに送信するステップ、
を実行させるシステム。
【請求項２】
請求項１記載のシステムであって、前記ローカルプライベートデータは、ローカルプライベート健康管理データを含むシステム。
【請求項３】
請求項２において、前記ローカルプライベート健康管理データは、患者固有のデータを含むシステム。
【請求項４】
請求項１において、前記ローカルプライベートデータは、ゲノムデータ、全ゲノム配列データ、全エキソソーム配列データ、プロテオームデータ、プロテオミクス経路データ、ｋ－ｍｅｒデータ、ネオエピトープデータ、ＲＮＡデータ、アレルギー情報、遭遇データ、治療データ、転帰データ、予約データ、注文データ、請求コードデータ、診断コードデータ、結果データ、治療反応データ、腫瘍反応データ、人口統計データ、投薬データ、バイタルサインデータ、支払者データ、薬物研究データ、薬物反応データ、経時的研究データ、バイオメトリックデータ、財務データ、所有権データ、電子カルテデータ、研究データ、人材データ、パフォーマンスデータ、分析結果データ、または事象データを含むデータの少なくとも１つを含むシステム。
【請求項５】
請求項１において、前記ネットワークは、無線ネットワーク、パケット交換ネットワーク、インターネット、イントラネット、仮想プライベートネットワーク、セルラネットワーク、アドホックネットワーク、およびピアツーピアネットワークの少なくとも１つを含むシステム。
【請求項６】
請求項１において、前記少なくとも１つの非プライベート演算デバイスは、前記トレーニング済み実モデルが作成されたローカルプライベートデータに対する権限がない前記複数のプライベートデータサーバのうちの異なる１つであるシステム。
【請求項７】
請求項１において、前記少なくとも１つの非プライベート演算デバイスは、グローバルモデルサーバを含むシステム。
【請求項８】
請求項７において、前記グローバルモデルサーバは、前記複数のプライベートデータサーバのうちの少なくとも２つからのプロキシデータのセットを集約するように構成され、グローバルモデルを前記プロキシデータのセットでトレーニングするように構成されているシステム。
【請求項９】
請求項１において、各前記プライベートデータサーバは、前記ローカルプライベートデータを格納するローカルストレージシステムに通信可能に接続されているシステム。
【請求項１０】
請求項９において、前記ローカルストレージシステムは、ＲＡＩＤシステム、ファイルサーバ、ネットワークアクセス可能なストレージデバイス、ストレージエリアネットワー

クデバイス、ローカルコンピュータ可読メモリ、ハードディスクドライブ、光ストレージデバイス、テープドライブ、テープライブラリ、およびソリッドステートディスクの少なくとも１つを含むシステム。

【請求項１１】

請求項９において、前記ローカルストレージシステムは、ローカルデータベース、ＢＡＭサーバ、ＳＡＭサーバ、ＧＡＲサーバ、ＢＡＭＢＡＭサーバ、および臨床オペレーティングシステムサーバの少なくとも１つを含むシステム。

【請求項１２】

請求項１において、前記モデル指示は、ローカルコマンド、リモートコマンド、実行可能ファイル、プロトコルコマンド、および選択されたコマンドの少なくとも１つを含むシステム。

【請求項１３】

請求項１において、前記複数のプライベートデータ分布の分布は、ガウス分布、ポアソン分布、ベルヌーイ分布、ラデマッハ分布、離散分布、二項分布、ゼータ分布、ガンマ分布、ベータ分布、およびヒストグラム分布の少なくとも１つに従うシステム。

【請求項１４】

請求項１において、前記複数のプライベートデータ分布は、前記トレーニング済み実モデルパラメータと前記ローカルプライベートデータから導出された固有値に基づくシステム。

【請求項１５】

請求項１において、前記プロキシデータのセットは、前記トレーニング済み実モデルパラメータと前記ローカルプライベートデータから導出された固有ベクトルの組み合わせを含むシステム。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 17/42356

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06N 99/00, G06N 5/04 (2017.01)
CPC - G06N3/126, G06N5/043, G06F15/18

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2014/0074760 A1 (Boldyrev et al.) 13 March 2014 (13.03.2014), entire document, especially para. [0030], [0041], [0044], [0049]	1-38
A	US 2011/0228976 A1 (Fitzgibbon et al.) 22 September 2011 (22.09.2011), entire document, especially [0075], [0076], [0131]	1-38
A	US 2012/0303558 A1 (Jaiswal) 29 November 2012 (29.11.2012), entire document, especially para. [0028], [0032], [0035], [0057]	1-38
A	US 2015/0170055 A1 (INT BUSINESS MACHINES CORP) 18 June 2015 (18.06.2015), entire document	1-38

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 September 2017	26 SEP 2017
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

フロントページの続き

(81)指定国・地域 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT

（特許庁注：以下のものは登録商標）
１．ＭＡＴＬＡＢ

(71)出願人 516001591
ナントミクス，エルエルシー
アメリカ合衆国，カリフォルニア州 ９０２３２，カルバー シティ，９９２０ ジェファーソン ブールバード
(74)代理人 110002572
特許業務法人平木国際特許事務所
(72)発明者 スゼト，クリストファー
アメリカ合衆国 ９５０６６ カリフォルニア州，スコッツ バレー，アルト ソル コート １２２
(72)発明者 ベンツ，スティーブン，チャールズ
アメリカ合衆国 ９５０６０ カリフォルニア州，サンタ クルーズ，スィート エー，ミッション ストリート ２９０１
(72)発明者 ウィッチェイ，ニコラス，ジェイ．
アメリカ合衆国 ９２６５３ カリフォルニア州，ラグーナ ヒルズ，ホン アベニュー ２４８３２

Ｆターム(参考) 5L099 AA21

【要約の続き】
【選択図】図１