

視覚と言葉を繋ぐAIのこれまでとこれから

オムロン サイニックエックス株式会社

  牛久 祥孝

  losnuevetoros

自己紹介

2013.6~2013.8	Microsoft Research Intern
2014.3	博士(情報理工学)、東京大学
2014.4~2016.3	NTT CS研 研究員
2016.4~2018.9	東京大学 講師 (原田牛久研究室)
2016.9~2018.9	産業技術総合研究所 協力研究員
2016.12~2018.9	国立国語研究所 共同研究員
2018.10~2024.3	オムロンサイニックエックス株式会社 Principal Investigator
2019.1~	株式会社 Ridge-i Chief Research Officer
2020.4~2023.3	津田塾大学 非常勤講師
2021.7~	東北大学 非常勤講師
2022.1~	合同会社ナインブルズ 代表
2023.12~	理化学研究所 客員主管研究員
2024.4~	オムロンサイニックエックス株式会社 VP for Research

画像キャプション生成

[Ushiku+, ACMMM 2012]

[Ushiku+, ICCV 2015]



A yellow train on the tracks near a train station.

動画の特定区間と キャプションの相互検索

[Yamaguchi+, ICCV 2017]



A guy is skiing with no shirt on and yellow snow pants.

2011

電話音声認識のエラー率が
30%程度→20%以下に

2012

大規模画像分類のエラー率が
25%程度→15%程度に

2014

英仏翻訳の精度が〇〇により
複雑なシステムと同等に

第3次AIブーム

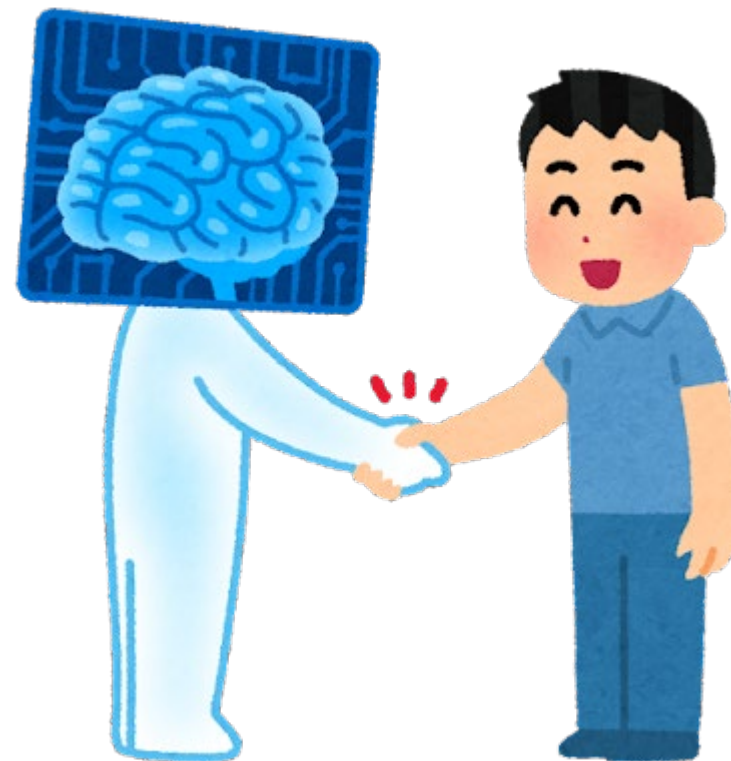
2022 生成AI元年

第4次AIブーム

と評する人もいる

今日のお話

- AIってなんだっけ？
- 視覚と言葉を繋ぐAIのこれまで ～最近はやりの生成AI
- 視覚と言葉を繋ぐAIのこれから ～AIロボット駆動科学



AIってなんだっけ？

最近のAIサービスの例 :

脳から、社会を変えていく。

カーボンニュートラルの実現

エネルギー可視化



カーボン・オフセット



デジタル化社会の実現

ロボット制御



画像分析+AI



アグリ自動化



vegetalia

オーガニック nice

メタバース



健康寿命の延伸

偏頭痛治療



認知症予防



医療データ



AST システム



遠隔医療



PCR 検査



デジタルコーチ



- 東北大発ベンチャー
 - MRI画像へのAI活用
 - 海馬の領域を自動検出
 - 海馬の体積を推定
- ➡認知機能低下リスクの評価

BrainSuite

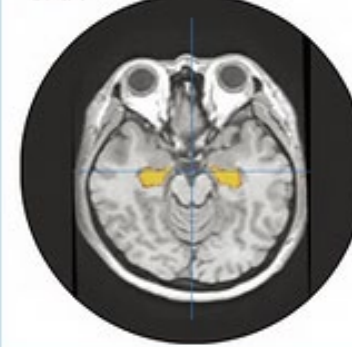
MRI撮影日	受診者ID	氏名	ページ
2022年4月15日	123456789012	海馬 小太郎 (かいば こたろう)	02/04

今回の海馬解析結果

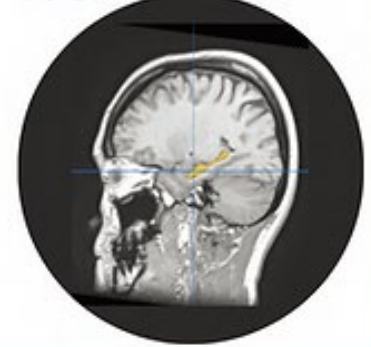
XXクリニックMRIによる海馬体積

7,350 mm³

横断面



矢状断面



海馬 (かいば)

記憶力をコントロールする脳領域の1つで、生活習慣等による変化が表れやすく、以下の特徴があるとされています。

✓認知機能の低下よりも先に萎縮が見られるため、早期の確認が重要

✓有酸素運動などにより神経細胞が新しく作られる「神経新生」が生じるため、萎縮のみならず維持・増大も期待できる



MyPage 脳を立体的に見る

MyPage (マイページ) では、ビューワー機能で上記のMRI画像を操作することで、自分自身の脳の様子を3Dで立体的に確認することができます。

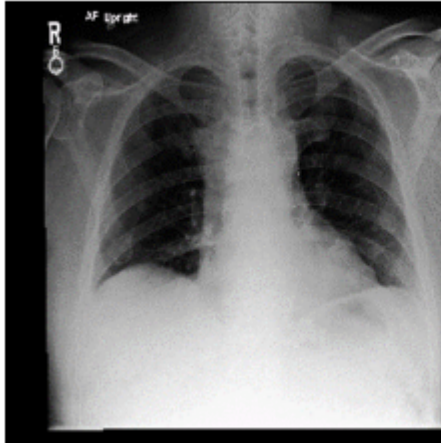
<https://mypage.brainsuite.jp/login>

最近のAI研究の例：X線読影レポート生成

[Li+, CVPR 2023]

X線画像

Image



医師の読影レポートの例

Ground Truth Report

lateral view somewhat limited due to overlying motion artifact lungs are low in volume there is no focal airspace the consolidation to suggest pneumonia a 1 2-cm calcified granuloma just below the medial aspect of the right hemidiaphragm is unchanged from prior study no pleural effusions or pulmonary edema there is no pneumothorax the inferior sternotomy wire is fractured but unchanged surgical clips and vascular markers in the thorax are related to prior cabg surgery.

AIが生成した例

Ours

lung volumes are low. there is no focal consolidation effusion or pneumothorax. airspace consolidation is noted within the left upper lobe compatible with pneumonia. borderline size of the cardiac silhouette without evidence of pulmonary edema. midline sternotomy wires and mediastinal clips are again noted.

(第3次ブーム以降の) AIってなんだっけ？

- 機械学習/深層学習によるAI
 - 深層学習は機械学習の一部
- 機械学習ってなんだっけ

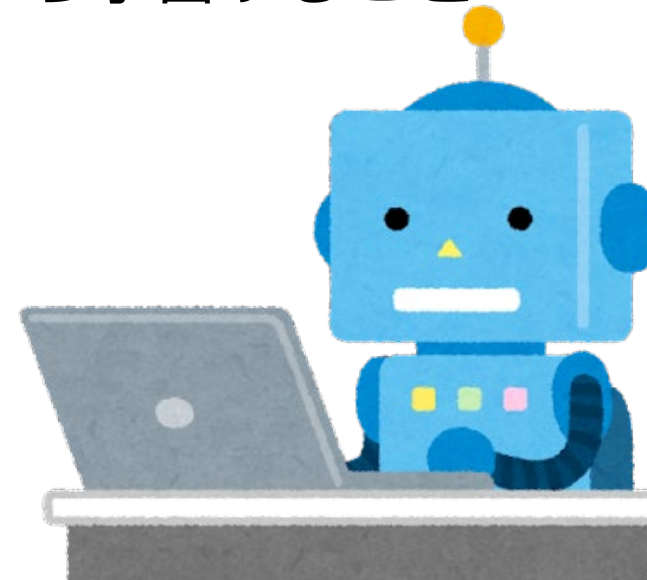
機械が関数

$$y = f(x)$$

↑ AIに出力させたいデータ

↑ AIに入力したいデータ

をデータから学習すること



2011

電話音声認識のエラー率が
30%程度→20%以下に

2012

大規模画像分類のエラー率が
25%程度→15%程度に

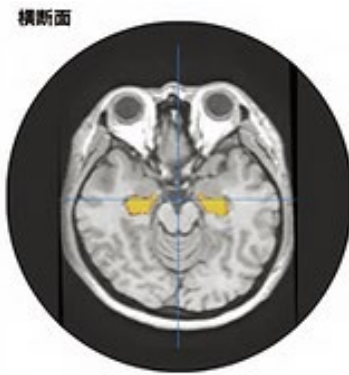
2014

英仏翻訳の精度が〇〇により
複雑なシステムと同等に

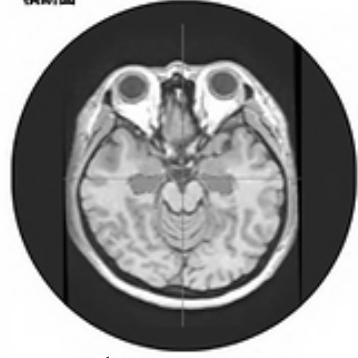
理解のAI

理解のAIその① 識別

入力データに対して離散的な値を出力するもの



=

f ()

それぞれの画素が
海馬かそれ以外かを識別

MRI画像
縦×横の画素の数値

理解のAIその② 回帰

入力データに対して連続的な値を出力するもの

$$7350 \text{ mm}^3 = f(\text{横断面}, \text{矢状断面})$$

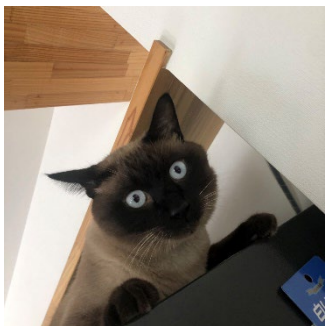
海馬の体積を回帰

上と横からみたMRI画像
縦×横の画素の数列

視覚と言葉を繋ぐAIのこれまで ～最近はやりの生成AI

深層学習前

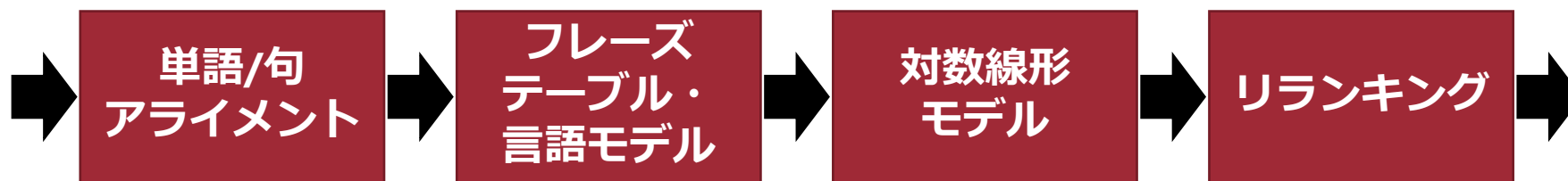
• 画像認識 (CV)



NLP研究者には謎

• 機械翻訳 (NLP)

吾輩は猫である。名前は
まだ無い。
どこで生れたかほとんど
見当がつかぬ。何でも薄
暗いじめじめした所で
ニャーニャー泣いていた
事だけは記憶している。



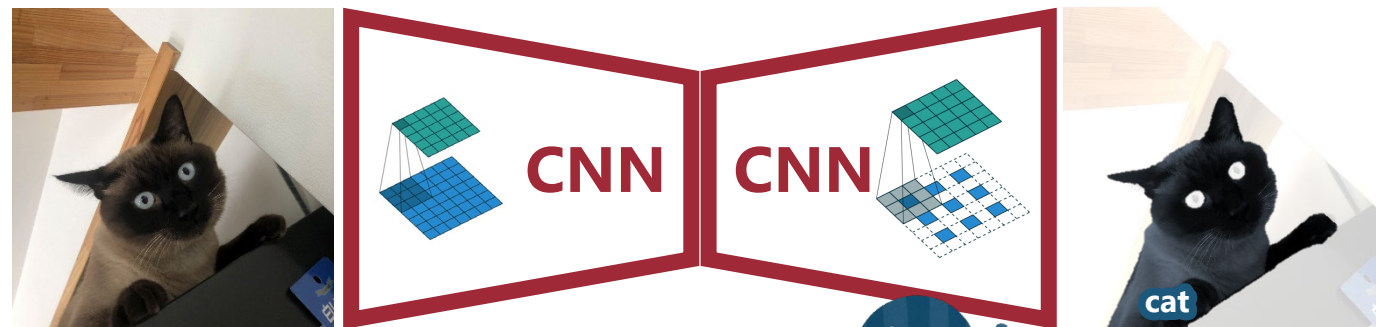
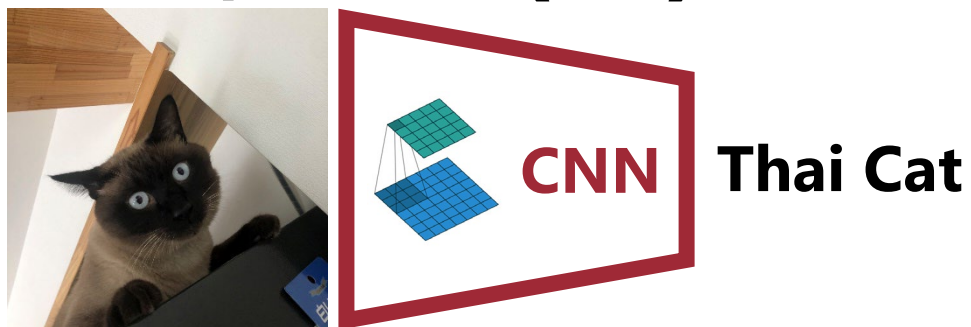
I am a cat. I don't
have a name yet.
I have no idea where I
was born. I only
remember that I was
crying in a damp and
dimly lit place.



CV研究者には謎

深層学習後

• 画像認識 (CV)



NLP研究者も... !

• 機械翻訳 (NLP)

吾輩は猫である。名前はまだ無い。
どこで生れたかほとんど見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。



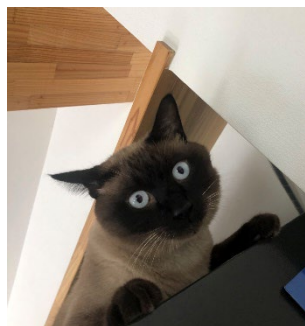
I am a cat. I don't have a name yet.
I have no idea where I was born. I only remember that I was crying in a damp and dimly lit place.



CV研究者も... !

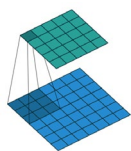
深層学習後

• 画像認識



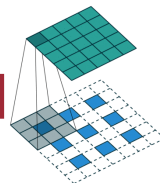
NLPでもCNNは使う！

吾輩は猫である。名前はまだ無い。
どこで生れたかとうとう見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。



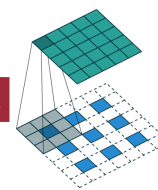
CNN

CNN



I am a cat. I don't have a name yet.
I have no idea where I was born. I only remember that I was crying in a damp and dimly lit place.

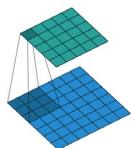
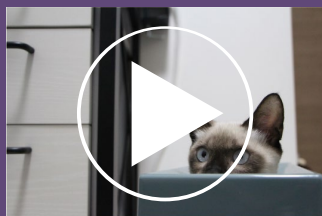
CNN



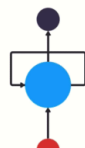
NLP研究者も...！

• 機械学習

CVでもRNNは使う！



CNN



RNN

Hiding



CV研究者も...！

ユーザー生成コンテンツの爆発的増加

特にコンテンツ投稿・共有サービスでは...

- Facebookにアップロードされる画像 **1日で計3億枚**
- YouTubeにアップロードされる動画 **1分間で計400時間分**



Pōhutukawa blooms this time of the year in New Zealand. As the flowers fall, the ground underneath the trees look spectacular.

画像/動画と
関連する文章の対
→大量に収集可能

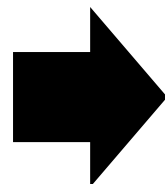
Vision and Language の萌芽的な研究

記事付き画像へのキャプション生成

[Feng+Lapata, ACL 2010]

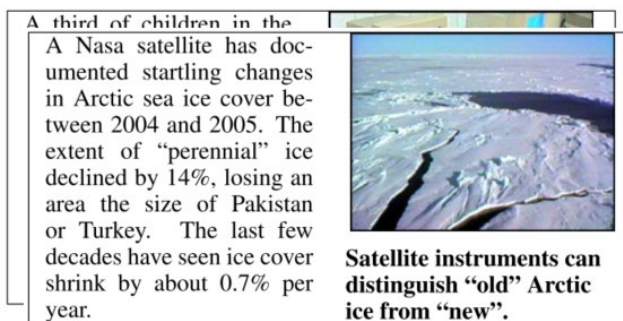
- Input: article + image Output: caption for image

Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. Representatives from 30 foreign countries watched as the king's coffin was carried by 1,000 men to the official royal burial ground.



King Toupu IV died at the age of 88 last week.

- Dataset: Sets of article + image + caption



× 3361

Vision and Language の萌芽的な研究

記事付き画像へのキャプション生成

[Feng+Lapata, ACL 2010]

- Input: article + image Output: caption for image

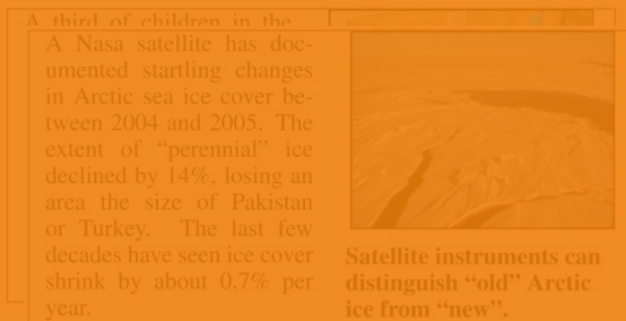
Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. Representatives from 30 foreign countries watched as the king's body was carried by palanquin to the official royal burial ground.



King Toupu IV died at the age of 88 last week.

これらの背景から...
**Vision&Languageと呼ばれる
様々な取り組みが！**

- Dataset: Sets of article + image + caption



Satellite instruments can distinguish "old" Arctic ice from "new".

× 3361

ビジュアル質問応答

[Fukui+, EMNLP 2016]



What vegetable is the dog
chewing on?

MCB: carrot

GT: carrot

ビジュアル質問応答

[Fukui+, EMNLP 2016]



What color is the traffic
light?

MCB: green

GT: green



①マルチモーダル理解

入力：ビジュアルデータ＋テキストデータ

→ 出力：認識結果

What color is the traffic
light?

MCB: green

GT: green

画像キャプション生成

[Ushiku+, ICCV 2015]



Group of people sitting at a table with a dinner.



Tourists are standing on the middle of a flat desert.

画像キャプション生成

[Ushiku+, ICCV 2015]



Group of people sitting
at a table with a dinner.

② Image2Text

入力：ビジュアルデータ

→ 出力：テキストデータ



Tourists are standing on
the middle of a flat desert.

キャプションからの画像生成

[Ramesh+, 2021]

An illustration of a baby hedgehog in a wizard hat riding a car



A photo of the food of japan



キャプションからの画像生成

[Ramesh+, 2021]

An illustration of a baby hedgehog in a wizard hat riding a car

A photo of the food of japan

③ Text2Image

入力：テキストデータ



出力：ビジュアルデータ



GPT-4V = Vision&Language版GPT

画像を用いた会話



この写真は地球上のどこで撮られたものを推定してください。



この写真は「Tokyo International Forum」という場所で撮られたものと思われます。この場所は日本の東京にあります。

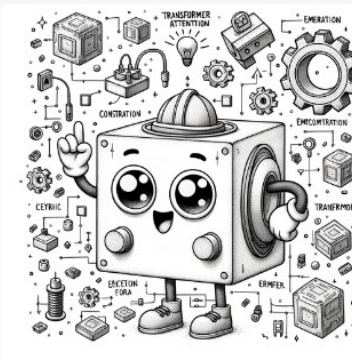
画像生成機能



あなたはプロのイラストレーターです。GPTという名前のかわいいキャラクターが、Transformerのattentionアーキテクチャを組み立てているイラストを制作してください。



DALL-E 3
作成された 4 枚の画像



Here are the illustrations of the cute character named GPT assembling the Transformer attention architecture. I hope you like them! If you have any feedback or would like further modifications, please let me know.

2022 生成AI元年

生成のAI

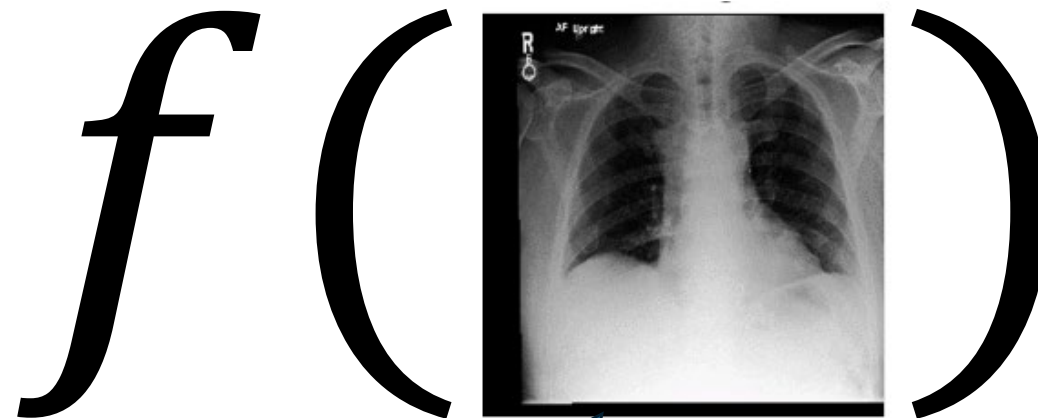
生成のAI

実は理解のAIと根本の技術は同じ

lung volumes are low. there is no focal consolidation effusion or pneumothorax. airspace consolidation is noted within the left upper lobe compatible with pneumonia. borderline size of the cardiac silhouette without evidence of pulmonary edema. midline sternotomy wires and mediastinal clips are again noted.

読影レポート

=



X線画像
縦×横の画素の数列

生成のAI

より具体的に説明すると

lung

=

$$f\left(\begin{array}{c} \text{X線画像} \\ \text{縦} \times \text{横の画素の数列} \end{array}\right)$$

最適な1単語目を識別

X線画像
縦×横の画素の数列

生成のAI

より具体的に説明すると

$$\text{volumes} = f\left(\text{X-ray image}, \text{lung}\right)$$

最適な2単語目を識別

X線画像と1単語目

生成のAI

より具体的に説明すると

$$\text{are} = f\left(\text{X-ray image}, \text{lung volumes}\right)$$

最適な3単語目を識別

X線画像と1~2単語目

生成のAI

より具体的に説明すると
➡自己回帰的な文生成

$$\text{low} = f\left(\text{X-ray image}, \text{lung volumes are}\right)$$

最適な4単語目を識別

X線画像と1~3単語目

ちなみに

画像生成AIの場合は



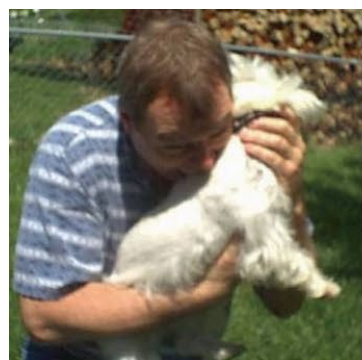
$= f(\text{あざといポーズをとる猫の絵を描いてください})$

カラー画像
縦×横の画素値を回帰

自然言語（テキスト）

牛久：もともとは画像キャプション生成芸人

[Ushiku+, ACM MM 2012][Ushiku+, ICCV 2015]



Keyphrases

his arms
man bites
white dog

+

Grammer

in → *his arms*
a → *man*
a → *white dog*

Output Sentence

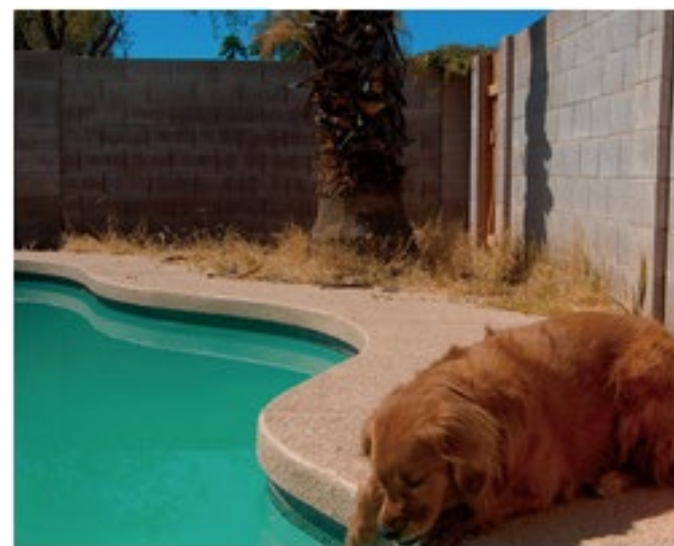
A man bites
a white dog
in his arms



A giraffe standing next to a tree in a fence.



A yellow train on the tracks near a train station.



A dog laying on the side of a zoo enclosure.

最近：Vision and Languageからの作業理解

[Nishimura+, MTA 2023]



- (a) Cut the pork in half and remove the pork
豚肉を半分に切り、豚肉を取り出す
- (b) Season the pork with salt and pepper
豚肉に塩・胡椒をする
- (c) Season the pork with salt and pepper
豚肉に塩・胡椒をする
- (d) Heat some butter in a pan
フライパンでバターを熱する
- (e) Coat the pork in the break crumbs
豚肉にパン粉をまぶします
- (f) Fry the pork in a pan
フライパンで豚肉を炒めます

[日経新聞2022年10月18日 朝刊

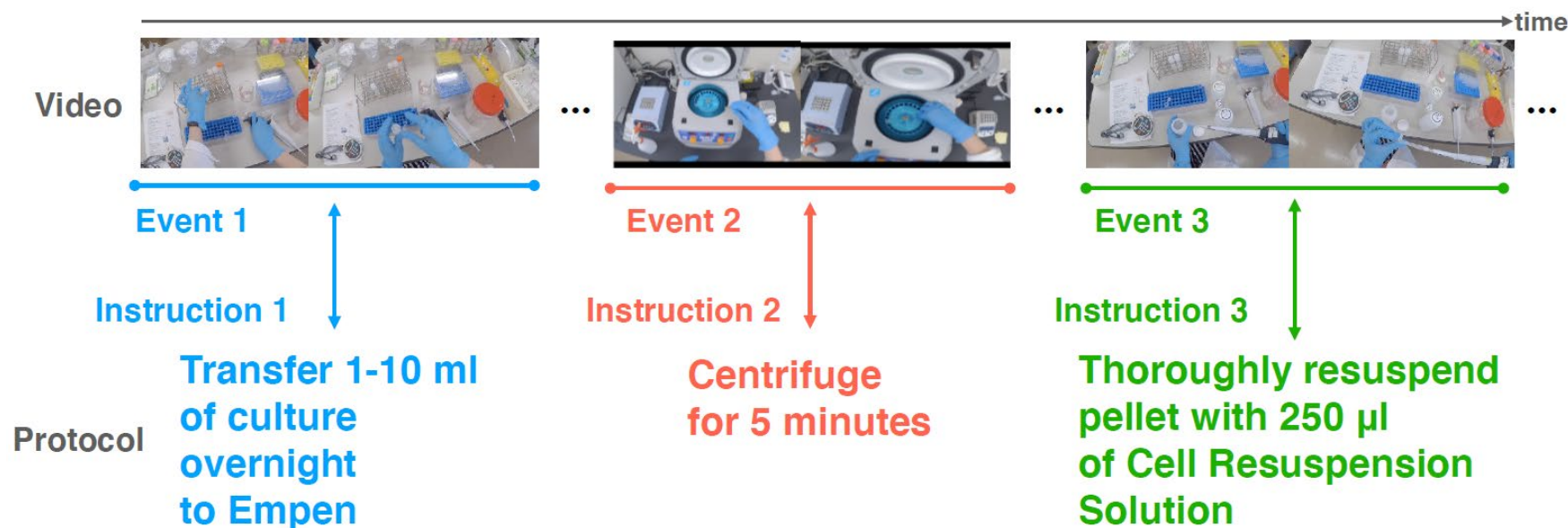
<https://www.nikkei.com/article/DGXZQOUC297B00Z20C22A9000000/>]



実験という作業を理解しよう→ロボットで補助しよう

[Nishimura+, ICCV WS 2021]

- 調理以外の作業を理解する→生化学実験
- 4種類の実験をそれぞれ4回撮影したデータセット
 - PCR, ミニプレップ, DNA抽出, アガノースゲル作成
 - DNA抽出は, フェノールクロロホルム抽出2回, エタノール抽出2回
 - 撮影後, 手順とvideoのイベントをアノテーション



視覚と言葉を繋ぐAIのこれから ～AIロボット駆動科学



YOSHUA BENGIO,
GEOFFREY E. HINTON
AND YANN LECUN

For conceptual and engineering
breakthroughs that have made
deep neural networks a critical
component of computing

[<https://awards.acm.org/about/turing-laureates-spotlight>]

A.M.
TURING

A W A R D
2018



遡ること20年前

YOSHUA BENGIO,
GEOFFREY E. HINTON
AND YANN LECUN

For conceptual and engineering
breakthroughs that have made
deep neural networks a critical
component of computing

[<https://awards.acm.org/about/turing-laureates-spotlight>]

A.M.
TURING

A W A R D
2018



A.M.
TURING
A W A R D
1998

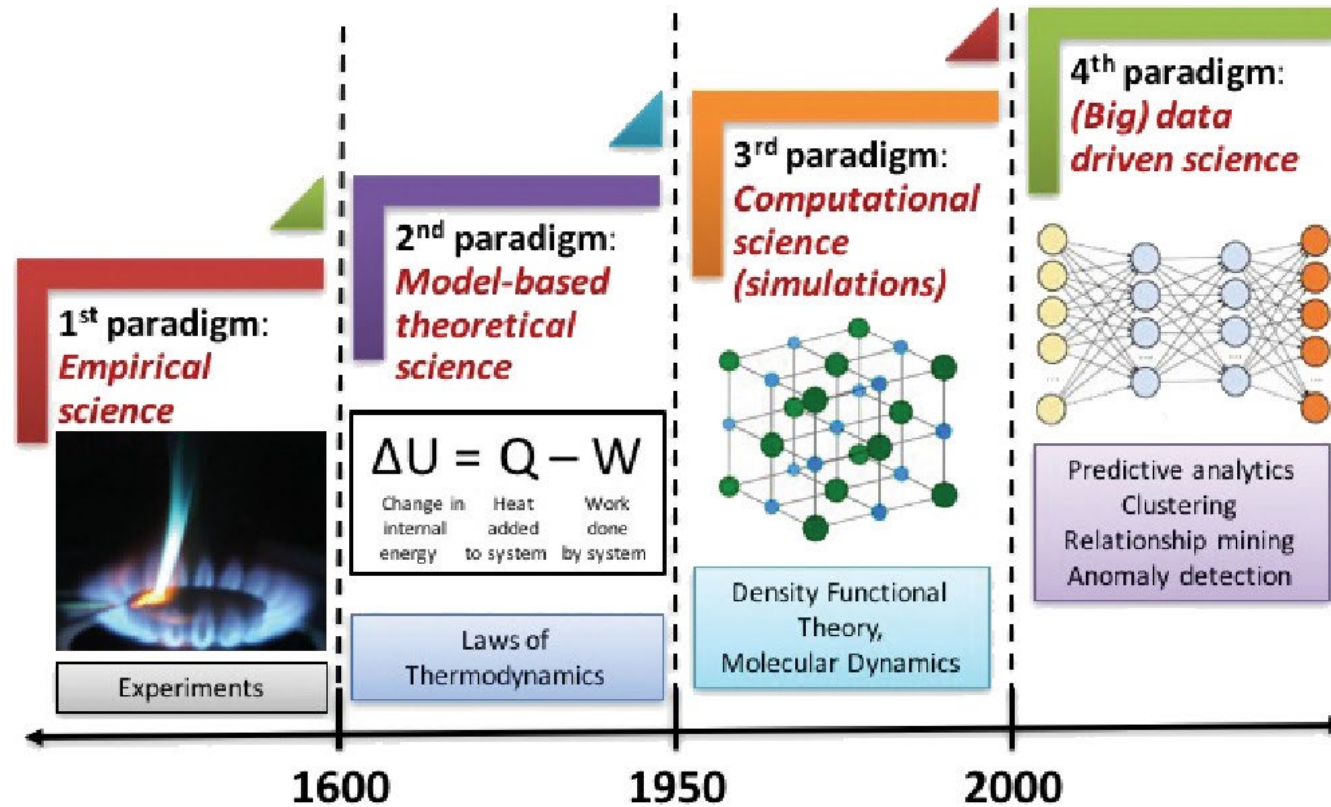


JIM GRAY

Transaction processing

[<https://awards.acm.org/about/turing-laureates-spotlight>]

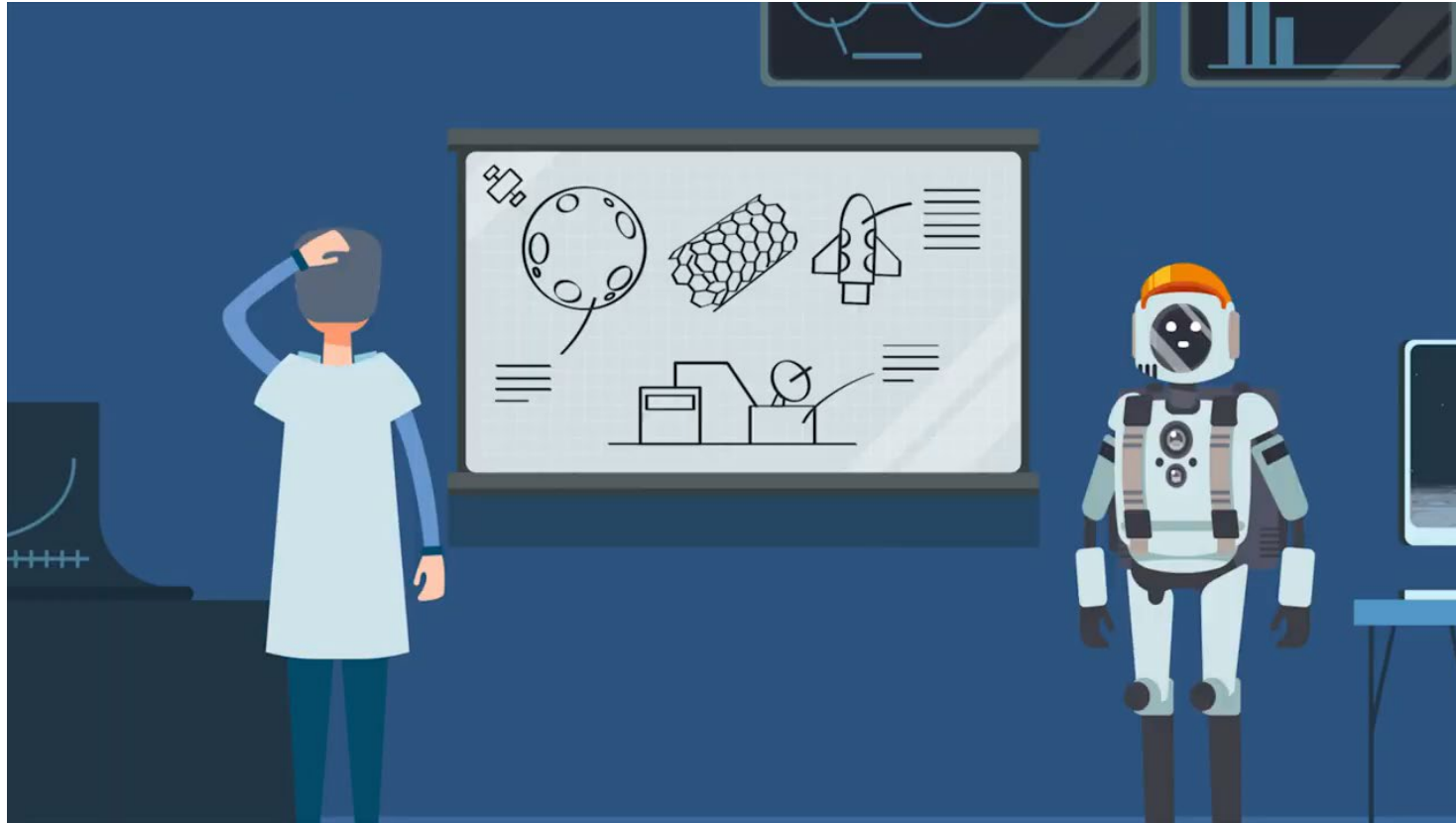
データドリブン科学の萌芽から20年



[Agrawal+Choudhary, APL Materials, 2016]

深層学習がデータドリブン科学を加速しているのが今！

人と融和して知の創造・越境をするAIロボット



2050



ノーベル賞級の研究を人とAIロボットが創出

やりたいこと：データから科学法則を発見したい



関数同定問題 Symbolic Regression

[Lalande+, NeurIPS 2023 WS][Matsubara+, DMLR'24]

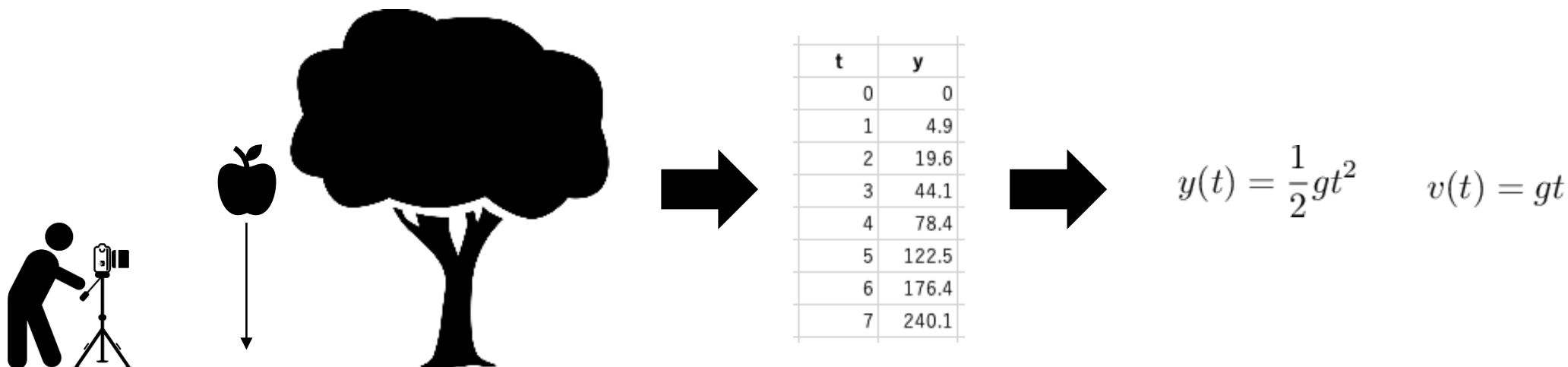
観測された実験結果を数式として理解する

- データセット：表

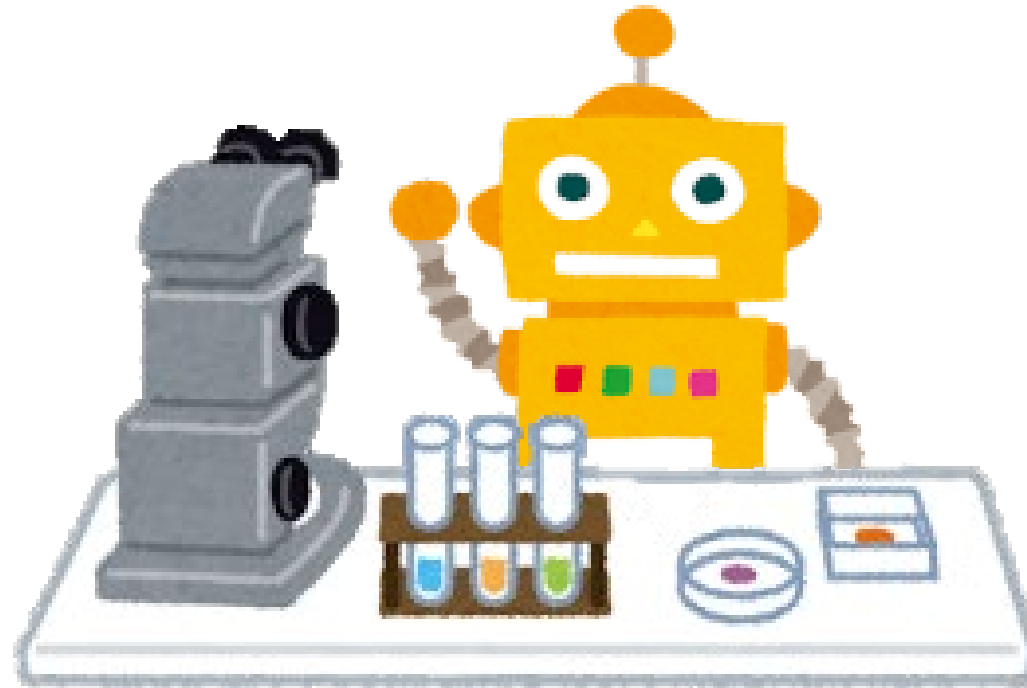
- 時刻
- リンゴのY座標

- 目的

- 変数間の関係を示す式の発見 = テーブルデータから数式への変換



やりたいこと：実験作業をロボットにやらせたい





まとめ

- AIってなんだっけ？
- 視覚と言葉を繋ぐAIのこれまで ~最近はやりの生成AI
- 視覚と言葉を繋ぐAIのこれから ~AIロボット駆動科学

人間の営みを理解し、
人間の創造的な活動をサポートするAIへ

