

医療・創薬データサイエンスコンソーシアム
H30年度イベント

Amazon Web Service機械学習PBLセミナー

神沼英里

東京医科歯科大学 医療データ科学推進室 特任講師

2018年11月7日（水） 13:00-18:00

AWS Japan オフィス(目黒雅叙園アルコタワーアネックス)

AWS機械学習PBLセミナーの目的

■目的：機械学習＋プログラミングのスキル向上

■インターンシップで求められる技能

＜某社データサイエンスの3カ月インターンシップ受入実績担当者＞

* データ収集に情報科学の知識

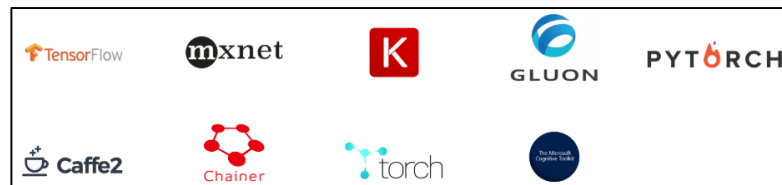
* データ解析にプログラミング能力

⇒ **上記2点の技能を持たない人物は、インターンシップに受入せず**

■インターンシップ代替イベントとしてデータサイエンスプログラミング実習を企画

* Project-Based Learning(PBL)型の
データサイエンス実践スキルアップセミナー開催(11/7)

AWS Japan社様の協力により、
Amazon Cloud(AWS)の機械学習ハンズオン実習



AWSのビルトイン深層学習ツール群



AWS機械学習PBLセミナー（11/7時間割）

時間	項目		準備事項
12:45	開場		* 出席確認 * AWSアカウントの受取 * アイデアソン番号の受取
13:00-14:00	第1部	Amazon SageMaker 概要の講義	
14:00-14:30	Q&A and Break		* AWSへログイン
14:30-16:00	第2部	機械学習モデル構築 実習	
16:00-16:30	Q&A and Break		* アイデアソンのチーム単位で着席 （ペア相手を探す） * 大判ポストイット1枚＋マーカー1 本を、チーム単位で受取
16:30-18:00	第3部	機械学習モデルの アイデアソン	
18:00	Close		

第3部 機械学習モデルのアイデアソン

第1回MD-DSC機械学習モデル アイデア賞 & モデリング賞

■応募内容

ライフサイエンス課題を解決する機械学習モデルの「アイデア」と「モデル」を募集する。

①アイデア (**アイデア賞**)

②機械学習モデル (**モデリング賞**)

本日アイデアソンの課題
= 機械学習モデルのアイデア出し

■機械学習モデルの適用分野

ライフサイエンス分野の課題を解決する、機械学習モデルとする。オープンデータに限られる為に、医療・創薬の分野に限らない。

■投稿フォーム



<https://tinyurl.com/ybn4sun4>

応募要項

- 応募内容：ライフサイエンス課題で機械学習モデルのアイデア（アイデア賞）、機械学習モデルの構築（モデリング賞）の提案を募集します。課題テーマは、ライフサイエンス分野とする（オープンデータが限られる為に、医療・創薬の分野に限定しない）。
- 応募締切：2019年1月8日
- 参加形態：個人orチームどちらでも可
- 参加資格：コンソーシアム受講生
- 提出物：アイデアや構築モデルの概要(PDFか画像)、解説ビデオ（モデリング賞のみ）下記を参照
- 審査方法
 - * MD-DSC研究会の1月開催日までに、あらかじめ受講生・参加企業で投票します
 - * 投票結果を元に、コンソーシアム運営会議で受賞者を決定します。
- 表彰式について
 - * 1月後半開催予定の第3回MD-DSC研究会で、上位入賞者を表彰します。
 - * 受賞チームは、表彰式でプレゼンテーションを行ないます。
 - * また当日、協賛企業から講評コメントを頂く予定です。
- 知財の取り扱いについて
 - * アイデア賞・モデリング賞の知的財産権は、医療・創薬データサイエンスコンソーシアムと受講生所属機関との個別相談になります。
- 注意事項
 - * 1人で複数回の応募が可能です
 - * 博士人材コース受講生で、アマゾン実習イベントとインターンシップのどちらにも参加していない方は、アイデア賞へ応募すれば修了要件を満たします

第3部 アイデアソン時間割

■ アイデアソンで4つの実習作業あり

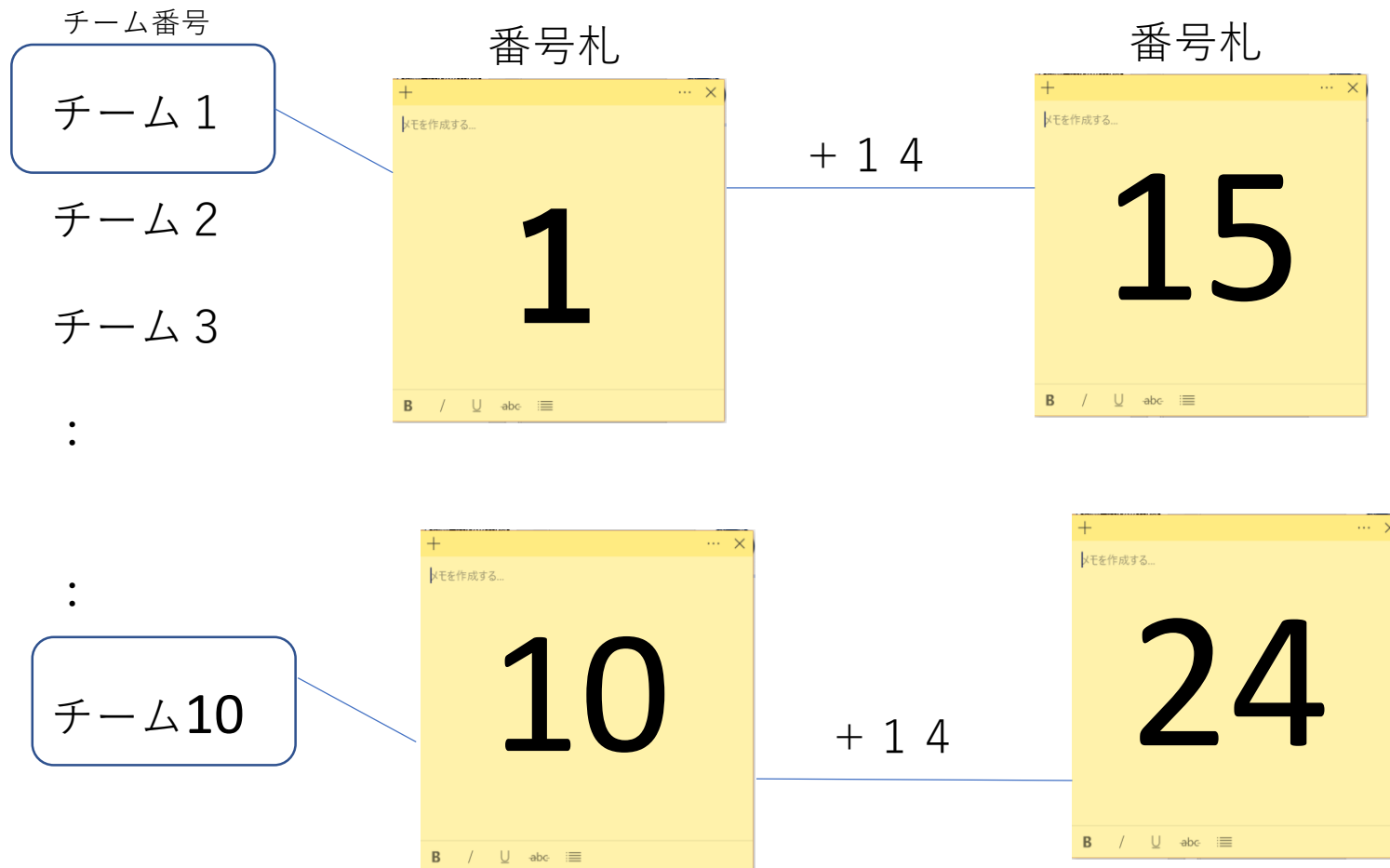
提案する機械学習モデルの①アイデアまとめ、②アイデア投稿、
③プレゼン準備、④プレゼン発表

■ 時間割

時間	項目	実習事項	物品
16:30-16:45	アイデアソン課題と説明	* アイデアソンのペア相手を探す 13チーム	
16:45-17:30	アイデア出し	* 機械学習モデルのアイデアを、 大判ポストイットにまとめる① * アイデア投稿② * 1分スピーチの準備③	* 大判ポストイット (1チーム：下書きと清書の2枚まで) * 記入用マーカー * 通常ポストイット
17:30-17:55	プレゼン発表 (1チーム1分)	* チーム毎に1分スピーチで アイデアを発表④	* ベル係、タイマー
17:55-18:00	まとめ		
18:00	Close		

※アイデア賞の評価は、第3回MD-DSC研究会(1月開催)で行います。

チーム構築（1チーム2名、別機関メンバ）



※ただし同一機関でチームになった場合は、
別機関メンバを探して番号札を交換して下さい。

機械学習モデルのアイデアの出し方①

データの種類起点で考える

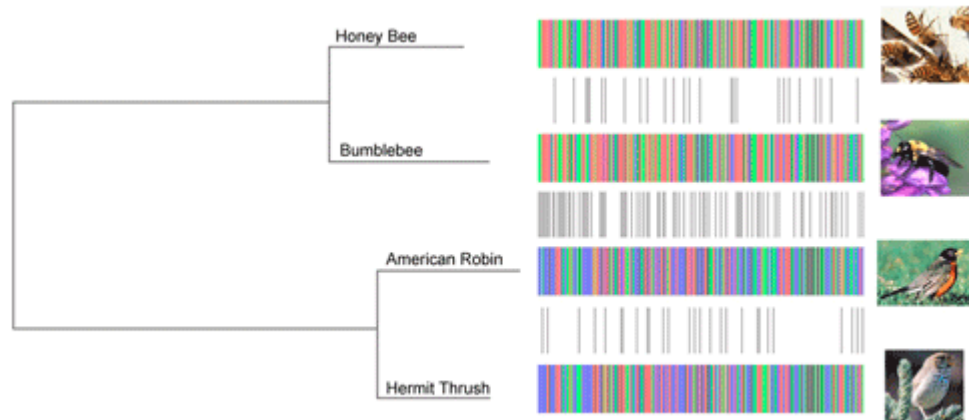
9

■ Computer Vision (画像分類など)



<https://www.maxpixel.net/>

■ DNA Sequence Analysis (DNA配列、オミックス情報の注釈など)



http://www.ecosmagazine.com/temp/EC11117_Fa.gif

■ 自然言語処理 (テキスト分類など)

Gene Chemical Disease Species Mutation Clear Reset

TITLE:
Cytotoxicity of Selenium Immunoconjugates against Triple Negative Breast Cancer Cells.

ABSTRACT:
Within the subtypes of breast cancer, those identified as triple negative for expression of estrogen receptor a (ESR1), progesterone receptor (PR) and human epidermal growth factor 2 (HER2), account for 10 20% of breast cancers, yet result in 30% of global breast cancer-associated deaths. Thus, it is critical to develop more targeted and efficacious therapies that also demonstrate less side effects. Selenium, an essential dietary supplement, is incorporated as selenocysteine (Sec) in vivo into human selenoproteins, some of which exist as anti-oxidant enzymes and are of importance to human health. Studies have also shown that selenium compounds hinder cancer cell growth and induce apoptosis in cancer cell culture models. The focus

PubTator

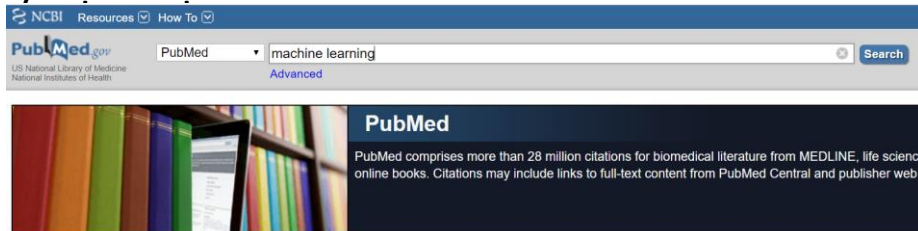
■ センサーデータ解析 (時系列解析など)



https://ja.wikipedia.org/wiki/FET09_Prague.jpg

①NCBI PUBMED

<https://www.ncbi.nlm.nih.gov>



- ☐ [Diabetic Retinopathy Diagnosis from Retinal Images Using Modified Hopfield Neural Network.](#)
- 34. Hemanth DJ, Anitha J, Son LH, Mittal M.
J Med Syst. 2018 Oct 31;42(12):247. doi: 10.1007/s10916-018-1111-6.
PMID: 30382410
[Similar articles](#)
- ☐ [EEG may serve as a biomarker in Huntington's disease using machine learning automatic classification.](#)
- 35. Odish OFF, Johnsen K, van Someren P, Roos RAC, van Dijk JG.
Sci Rep. 2018 Oct 31;8(1):16090. doi: 10.1038/s41598-018-34269-y.
PMID: 30382138 **Free Article**
[Similar articles](#)
- ☐ [Artificial Intelligence and amniotic fluid multiomics analysis: The prediction of perinatal outcome in asymptomatic short cervix.](#)
- 36. Bahado-Singh RO, Sonek J, McKenna D, Cool D, Aydas B, Turkoglu O, Bjorndahl T, Mandal R, Wishart D, Friedman P, Graham SF, Yilmaz A.
Ultrasound Obstet Gynecol. 2018 Oct 31. doi: 10.1002/uog.20168. [Epub ahead of print]
PMID: 30381856
[Similar articles](#)
- ☐ [Latent Factors and Dynamics in Motor Cortex and Their Application to Brain-Machine Interfaces.](#)
- 37. Pandarinath C, Ames KC, Russo AA, Farshchian A, Miller LE, Dyer EL, Kao JC.
J Neurosci. 2018 Oct 31;38(44):9390-9401. doi: 10.1523/JNEUROSCI.1669-18.2018.
PMID: 30381431
[Similar articles](#)

②米コーネル大 arXiv

<https://arxiv.org/>

- [1] [arXiv:1811.00995 \[pdf, other\]](#)
Invertible Residual Networks
Jens Behrmann, David Duvenaud, Jörn-Henrik Jacobsen
Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV); Machine Learning (stat.ML)
- [2] [arXiv:1811.00986 \[pdf, other\]](#)
Anomaly Detection for imbalanced datasets with Deep Generative Models
Nazly Rocío Santos Buitrago (1), Loek Tonnaer (1), Vlado Menkovski (1), Dimitrios Mavroidis (2) ((1) Eindhoven University of Technology, Royal Philips B.V., Eindhoven, The Netherlands)
Comments: 15 pages, 13 figures, accepted by Benelx 2018 conference
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
- [3] [arXiv:1811.00972 \[pdf, other\]](#)
Clustering and Learning from Imbalanced Data
Naman D. Singh
Comments: 9 pages, Submitted to NIPS 2018 Workshops
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
- [4] [arXiv:1811.00958 \[pdf, other\]](#)
Dantzig Selector with an Approximately Optimal Denoising Matrix and its Application to Reinforcement Learning
Bo Liu, Luwan Zhang, Ji Liu
Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Machine Learning (stat.ML)

③arXiv日本語訳有志 arXivTimes

<https://arxivtimes.herokuapp.com/>

Computer Vision

Recent Quality

-  **Graph Convolutional Reinforcement Learning for Multi-Agent Cooperation**
Graph Convolutionを利用したマルチエージェントの強化学習を解く手法の提案。各エージェントの観測結果をノードとし、Graph Convolutionをかけた結果を観測情報と併せて各エージェントのQ-Networkに入力する。ただ、接続関係(Edge)をどう定義するかは環境に依存する。
score:67  [arXiv:1811.00947](#) 2018-10-31  **ComputerVision** **ReinforcementLearning**
-  **Knows When it Doesn't Know: Deep Abstaining Classifiers**
学習データに含まれるノイズの影響を軽減する手法。分類クラスに「不要クラス」を一つ追加し、イメージ的にはモデルがそこにデータを「捨てる(Abstain)」ことを許容する形で学習を行う。既存のsoftmaxに不要クラス確率を組み込んだ式+捨てすぎ抑制の項というlossで学習する
score:50  [arXiv:1811.00947](#) 2018-10-25  **ComputerVision** **Optimization**
-  **Do Deep Generative Models Know What They Don't Know?**
生成モデルは入力データの分布をモデル化するため、外れ値(out-of-distribution)への対応にも強いと考えられていた。ところが、実験してみると学習したデータよりも学習していないデータに対し高い尤度を割り当てた現象が見られた(CIFAR-10/SVHNで確認)。ただ、これが一般的な現象なのかは要検証。
score:46  [arXiv:1811.00947](#) 2018-10-23  **ComputerVision** **Generation**
-  **Unconventional Wisdom: A New Transfer Learning Approach Applied to Bengali Numeral Classification**
CNNを転移学習する際、最初と最後の層「以外」をfreezeするという通常とは変わった手法の提案。最終の分類層をランダムな重みで固定してもなかなかの精度が出るという。Kaggleで開催されたベンガル語手書き数字を認識するコンペティションで6位をとっている。なお上位はより大きいモデル+アンサンブル

①AWS上のオープンデータ

<https://registry.opendata.aws/usage-examples/>

Search datasets (currently 15 matching datasets)

life sciences

Allen Brain Observatory - Visual Coding AWS Public Data Set

neurobiology neuro imaging image processing machine learning life sciences

The Allen Brain Observatory – Visual Coding is the first standardized in vivo survey of physiological activity in the mouse visual cortex, featuring representations of visually evoked calcium responses from GCaMP6-expressing neurons in selected cortical layers, visual areas, and Cre lines.

[Details →](#)

Usage examples

- [Use the Allen Brain Observatory – Visual Coding on AWS](#) by Nika Keller, David Feng

[See 1 usage example →](#)

TCGA on AWS

cancer genomic life sciences

The Cancer Genome Atlas (TCGA) is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to accelerate our understanding of the molecular basis of cancer. TCGA-funded researchers across the United States have produced a corpus of raw and processed genomic, transcriptomic, and epigenomic data from thousands of cancer patients.

[Details →](#)

Usage examples

- [AWS Building High-Throughput Genomics Batch Workflows on AWS](#) by Aaron Friedman

[See 1 usage example →](#)

②日本政府のオープンデータ

<http://www.data.go.jp/>



ホーム / データセット

組織

厚生労働省 (1931)

データセットを検索...

メタデータダウンロード

関連性

降順

1,931 件のデータセットが見つかりました

組織: 厚生労働省

11月27日の暴風雪による災害救助法の適用について

11月27日の暴風雪による災害救助法の適用についてのプレスリリース。

PDF

リリース日: 2012-11-27

メタデータ更新日: 2018-09-21

厚生労働省防災業務計画

厚生労働省の所掌事務について、防災に関し講ずるべき措置及び地域防災計画の作成の基準となるべき事項等を定めて防災行政事務の総合かつ計画的な遂行に資することを目的とする計画である。

HTML PDF

リリース日: 2017-07

メタデータ更新日: 2018-09-21

平成27年度 給水人口と水道普及率

このデータセットには説明がありません

PDF

リリース日: 2013-03-31

メタデータ更新日: 2018-09-21

平成27年度 水道の種類別箇所数

このデータセットには説明がありません

実践①アイデアをまとめよう

- [1] 大判ポストイット 1 枚 + マーカをチームで受取
(下書き用1枚有)
- [2] 自己紹介 + チーム名の決定
- [3] 機械学習モデルのアイデアを、大判ポストイットにまとめる
 - * 対象分野はライフサイエンスに限定
 - * アイデアの場合は、アルゴリズムは評価できない

■ ライフサイエンス
以外の事例

参考モデル

AstroNet

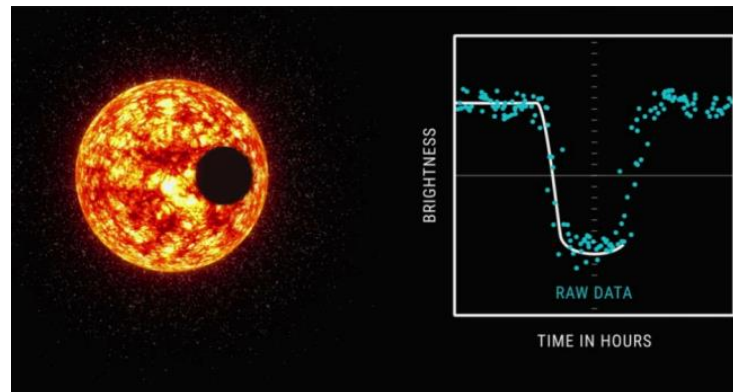
(NASA, Google Brain)

機械学習モデル名：**惑星か否かの判定モデル**

背景・問題：惑星画像は量が膨大でノイズが多い

提案学習データ：NASAのKepler宇宙望遠鏡の画像

提案アルゴリズム：畳み込みニューラルネットワーク



<https://github.com/tensorflow/models/tree/master/research/astronet>



賞の区分 *

☒ アイデア賞

☐ モデリング賞

機械学習モデル名 *

惑星か否かの判定モデル

チーム名 *

個人の場合は、お名前を入れて下さい。

目黒駅徒歩8分チーム

チーム参加者名 (全員) *


全員の名前 (所属先) をカンマ区切り(,)で入力して下さい。先頭はチーム代表者にして下さい。コンソーシアム外部メンバーが入る場合は、チーム代表者名 (所属先)、外部メンバー名 (所属先、外部)、以下略、と記載して下さい。

参加者A(***会社)、参加者B(***大学)

機械学習モデルの1行説明 *

惑星か否かを分類する深層学習モデル。NASA Kepler宇宙望遠鏡の画像を、モデルの学習に使用する。

アイデア・モデルの概略(PDF or 画像、A4サイズ1~2枚)を投稿して下さい *

 SampleAstroNet.P... X

アイデアまとめ (大判ポストイット) を写真撮影した画像をアップロード

解説ビデオURL (モデリング賞のみ)

JupyterNotebookの画面上で、機械学習モデルのデータ読み込みから推論までを解説したビデオ(MP4形式推奨)を作成して下さい。ビデオのダウンロード先URLを、ここに記載して下さい。オンラインにビデオファイルを掲載できない場合は、USBメモリに入れて直接1号館2階の学長戦略企画課に提出して下さい。

回答を入力

指導教員・所属長への確認 *

指導教官等に、アイデア賞&モデリング賞への応募について、連絡する必要がある場合は確認をお願い致します。投稿内容についての知的財産が生じる場合は、表彰式後に、医療・創薬データサイエンスコンソーシアムと受講生の所属機関との間で知財内容を詰める流れになります。

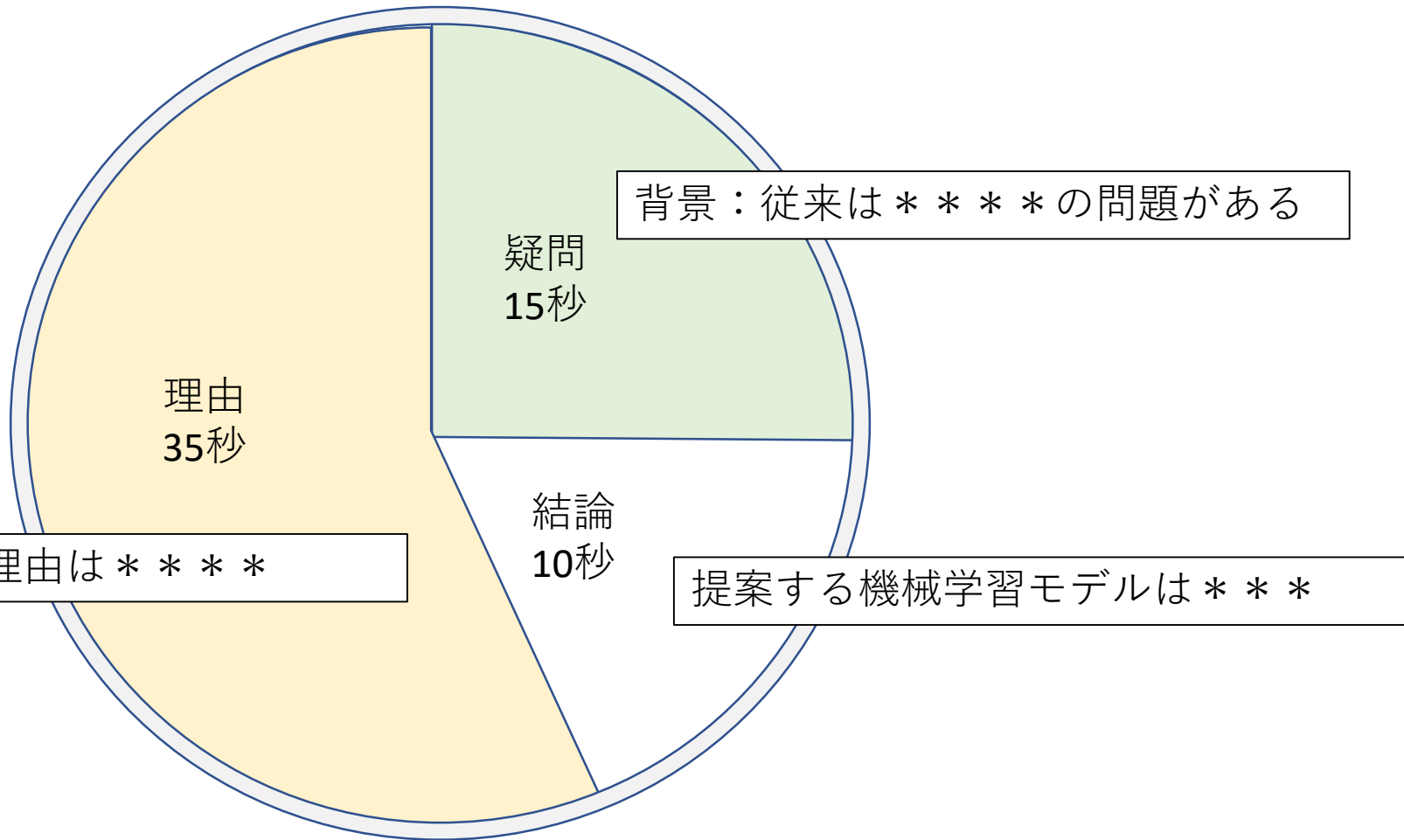
☒ 指導教員などへの確認について承知しました

Googleアカウントが無く、アイデアまとめ画像がアップロードできない場合
①空欄のままアイデアを投稿
②画像は下記宛で送信 (件名にチーム名を入力)

メール送付先: ekmds@tmd.ac.jp

③ 1 分間スピーチを準備しよう

参考情報＝シナリオ構成は、疑問15秒→結論10秒→理由35秒



④ 1 分間スピーチの手順

[1] 全員が見える場所(壁 etc)に、大判ポストイットを貼付。

[2] アイデア賞の投稿順に、コーディネータがチーム名を呼ぶ

[3] チーム名とメンバーの紹介

[4] アイデア1分発表



[5] 次チームに交代



第1回MD-DSC機械学習モデル アイデア賞&モデリング賞は、2019年1月8日まで募集中

16

※アイデアソン投稿作品も含む

※何件でも応募可能

※チームに限らず、個人でも応募可能

※第3回MD-DSC研究会で、上位アイデア & モデルを表彰予定

モデリング賞にもチャレンジしてみよう！！！！