

医療・創薬データサイエンスコンソーシアム
2019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

2019年11月28日 (木) 13:00-18:00

AWS Japan オフィス(目黒セントラルスクエア 17F)

AWS機械学習PBLセミナー（19/11/28時間割）

■目的：機械学習プログラミングの実践スキル習得

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2回MD-DSC機械学習コンペティション (アイデア賞&インプリメント賞)

■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(**インプリメント賞**)

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。オープンデータが限られる為に、医療・創薬の分野に限らない。

■投稿フォーム



<https://tinyurl.com/ybn4sun4>

昨年度応募作品：アイデア賞に16件、インプリメント賞1件

ID	機械学習モデル名	機械学習モデルの1行説明	チーム名
A1	子供の肥満に関連する妊娠前の母親の食事パターンの解析	コホート研究で食事の撮影画像を収集し、食事のパターン認識を実施	栄養解析チーム
A2	赤ちゃんの泣き声翻訳機 (アイデア賞1位)	赤ちゃんの泣き声からしてほしいケアが他人でもわかる	銀だこSS
A3	ギャンブル依存症になりそうかどうかの判定	ギャンブル依存症になりそうかどうかの判定モデル	チーム418
A4	がんゲノム情報翻訳モデル	がんゲノムの変異データをデータベースから適切な情報を張り付けていくと同時にデータベースの英語情報を翻訳してレポート化する翻訳機械学習モデル。	スキー好き
A5	運転中突発事故判定モデル (特別賞)	運転中に突発的な事故を起こすかどうかの判定モデル	チーム418
A6	病院努力の促進モデル	病院努力を促すために、病院のヒト・質・モノと治療成績などをもとに、病院の評価を客観的に行います。	ひさやま
A7	未知の副作用の予測	未知の副作用を予測する機械学習モデル。PMDAのJADERの副作用データを、モデルの学習に使用する。	信濃町
A8	自然災害時のより効率的なレスキュー部隊、救援物資の、輸送方法の予測モデル (アイデア賞3位)	自然災害時の効率的なレスキュー方法予測モデル	チームレスキュー

昨年度応募作品:アイデア賞に16件、インプリメント賞1件(続)

A9	「病は天気から？」予測	気候データ、診断情報から病気や事故などの発症、発生の予測を行う。	はぐれ者
A10	ePROに用いるための服薬確認用画像認識アプリ	ePROに用いることを前提とした服薬確認用画像認識アプリです。	M&S
A11	健康か否かの判定モデル	顔写真から健康かどうかを判定するモデル	早生まれ
A12	NLPを用いたGene Signatureの Interpretation	既存の方法では、有効な解析を行うことが困難であった網羅的解析に対し、NLPにより、意味付けを行うことができる。	アフターX'mas
A13	毒きのこ判別モデル (ピアレビュー賞、特別賞)	毒きのこを瞬時に食べられるかを判断できるモデル	グループ1
A14	認知症早期発見モデル	MRI画像などの医療データと、日常活動の変化(趣味の変化など)などのマルチモーダル情報学習	チーム岐阜
A15	AI 吉夢 (きちむ) システム	脳波をモニターする人工知能システムによって、寝たきり高齢者の良質な睡眠を確保する	安田永智
A16	仮面様顔貌の素顔を見抜くモデル (アイデア賞2位)	見過ごされがちだったパーキンソン病の一症状である仮面様顔貌に焦点を当て、顔写真画像データから病型診断・重症度評価を可能にするモデル	吉岡 耕太郎
M1	アンサンブル学習による認知症患者の識別タスク (特別賞)	認知機能テスト(MMSEのスコア)および脳MRI検査から得られる特徴量(脳容積の計算値)のOASIS(オープンアクセス) データを利用し、アンサンブル学習の正解率(78.5%)が、サポートベクターマシン、決定木、XGBoost のそれぞれ単体の学習器の正解率を数パーセント、上回ることを確認した。	安田永智



今年度：第2回MD-DSC機械学習コンペティション (アイデア賞 & インプリメント賞)

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）、または実装プログラム（インプリメント賞）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（オープンデータに限られる為に、医療・創薬の分野に限定しない）。
- 応募締切： **2020年1月6日(月) 9:00**
- 参加形態：個人orチームどちらでも可
- 参加資格： **今年度と昨年度のコンソーシアム受講生(2018年度、2019年度)**
- 提出物：アイデアや実装モデルの概要(PDFか画像)
- 審査方法
 - * 1月9日までに、あらかじめ連携企業審査員が投票します
 - * 投票結果を元に、コンソーシアムの連携機関連絡会議で受賞者を決定します。
- 表彰式について
 - * 1月20日（月）午後開催予定の第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・インプリメント賞の知的財産権は、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
 - * **インプリメント賞への応募により「応用編の修了課題」とします。**

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）

チーム番号

チーム 1

チーム 2

チーム 3

:

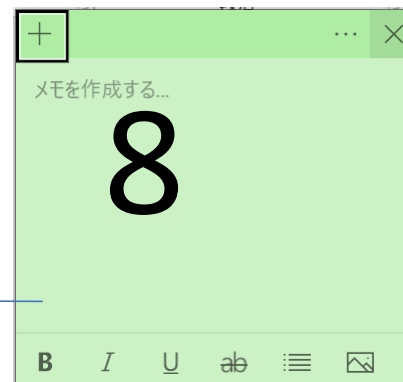
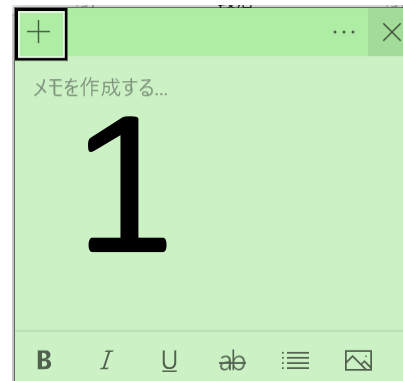
:

チーム 8

番号札(企業人材)



番号札（博士人材）



※端数の場合は、3名で1チームになります。

機械学習モデルのアイデアの出し方①

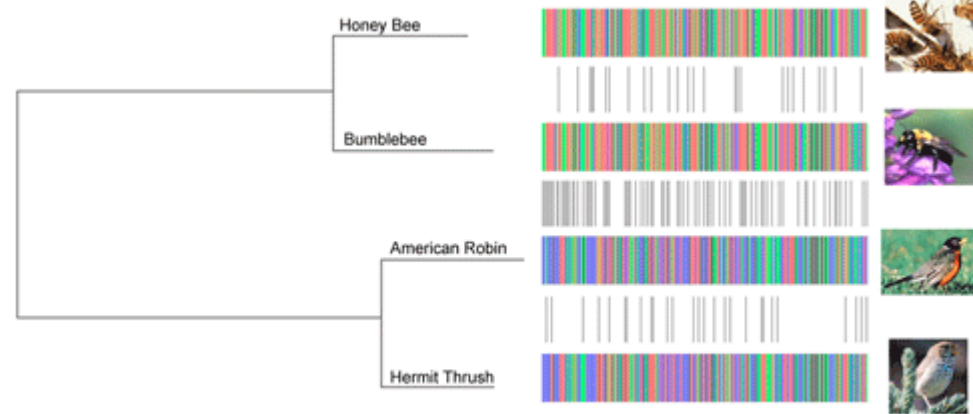
データの種類起点で考える

■ 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

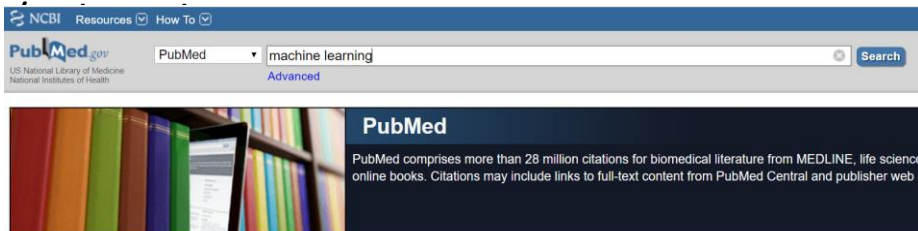
■ センサーデータ解析 (時系列解析など)



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

① NCBI 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, (2) Royal Philips B.V., Eindhoven, The Netherlands)
Comments: 15 pages, 13 figures, accepted by Benelearn 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)

③ arXiv 日本語訳有志 arXivTimes

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

Computer Vision

- ☒ [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 [icofog417](#) 2018-10-31 **CNN** 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 [icofog417](#) 2018-10-25 **CNN** 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 [icofog417](#) 2018-10-23 **CNN** 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的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



表彰式当日に、審査委員の選考以外にも、ピアレビュー賞をシール投票で決めます。



賞の区分 *

☒ アイデア賞

☐ モデリング賞

機械学習モデル名 *

チーム名 *

個人の場合は、お名前を入れて下さい。

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。

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... ✕

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

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）

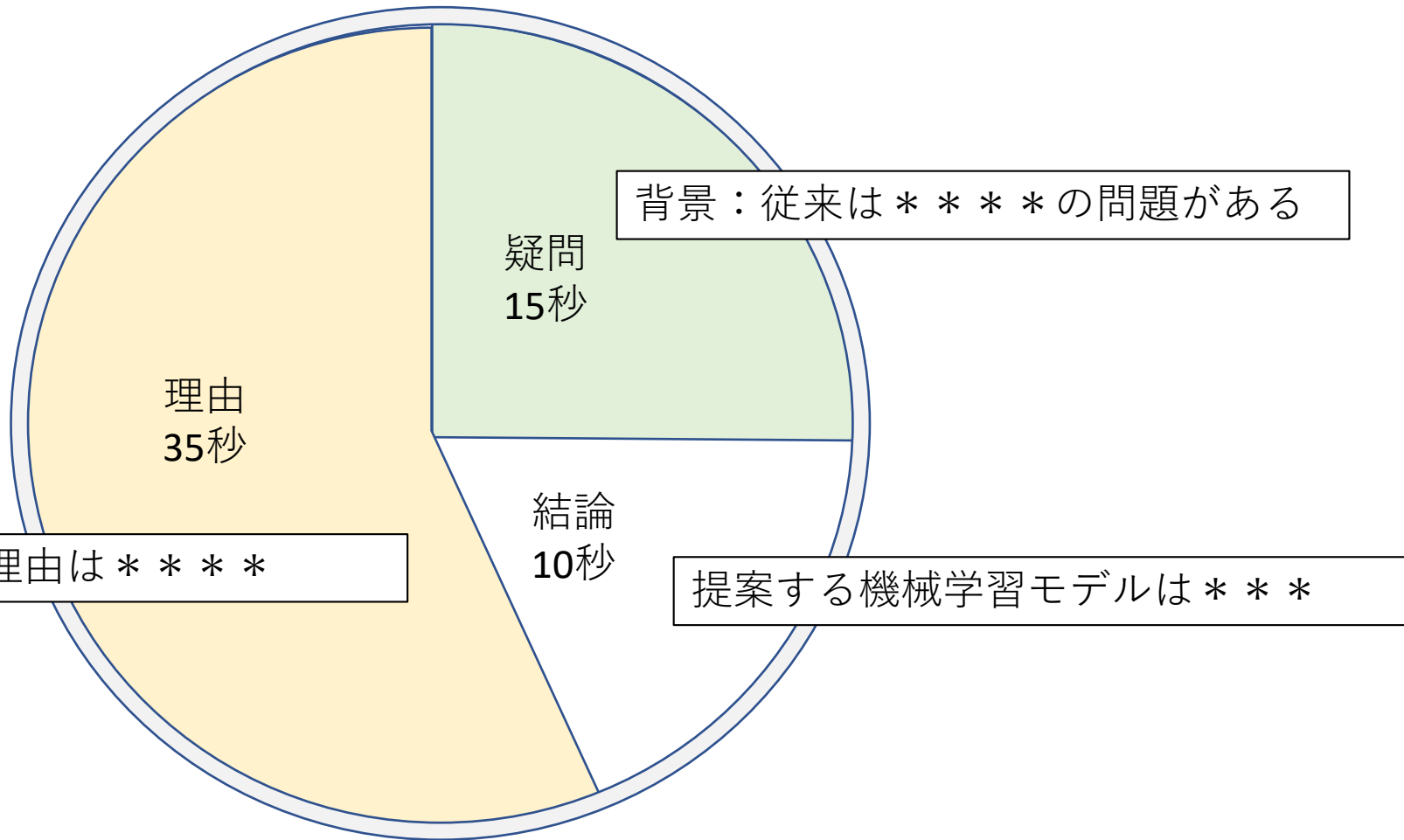
メール送付先: ekmds@tmd.ac.jp

BACKUP



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、2020年1月6日9時まで募集中

※学習データはこれから収集予定でOK（アイデア賞）

※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(1月20日)で、上位アイデア & 実装モデルを表彰予定

インプリメント賞にもチャレンジしてみよう！！！！