DATA｜データを活用した教育の実践
教育デザインと情報メディアを考えるシンポジウム 2019

データ分析コンテストとデータサイエンティストの働きかた

2019/12/14

株式会社ディー・エヌ・エー
AIシステム部 データサイエンス2G
グループマネージャー 松井 健一

:DeNA Delight and Impact the World

本日の講演の概要

データ分析に普段接していない方にも理解いただける
ことをターゲットにしています。

データ
サイエンティスト

専門

データ分析

活用

ビジネス

- ✓可視化、レポーティング
- ✓統計的推測、統計モデリング
- ✓**予測モデル構築**
- ✓エンジニアリング・プログラミング
- ✓ドメイン知識

参加

機械学習

データ分析コンテスト

Kaggle

アジェンダ

1. 自己紹介
2. データサイエンティストとは？
3. 機械学習とは？
4. データ分析コンテストの仕組み
5. 事業で活躍するKaggler
6. おわりに

自己紹介：松井健一　(Ken’ichi Matsui)

株式会社 ディー・エヌ・エー AIシステム部
データサイエンス第２グループ　グループマネージャー

経歴

大手SIer⇒大手通信キャリア⇒外資系コンサルティングファーム⇒現職

主な職務経験

- DRIVE CHARTににおける危険シーン検知の開発
- センサーデータによる異常検知
- 地理情報解析による人口動態分析
- Deep Learning/機械学習による画像解析
- SNS解析による商品評判解析
- ドライブデータの解析 (DRIVE CHART)
- 「アクセンチュアのプロフェッショナルが教える データ・アナリティクス実践講座」共著

https://drive-chart.com/

ブログ (Qiita: 技術系ブログサイト)

2015年～2016年頃は統計、機械学習、プログラミングに関するブログをよく書いていました。

DeNA

アジェンダ

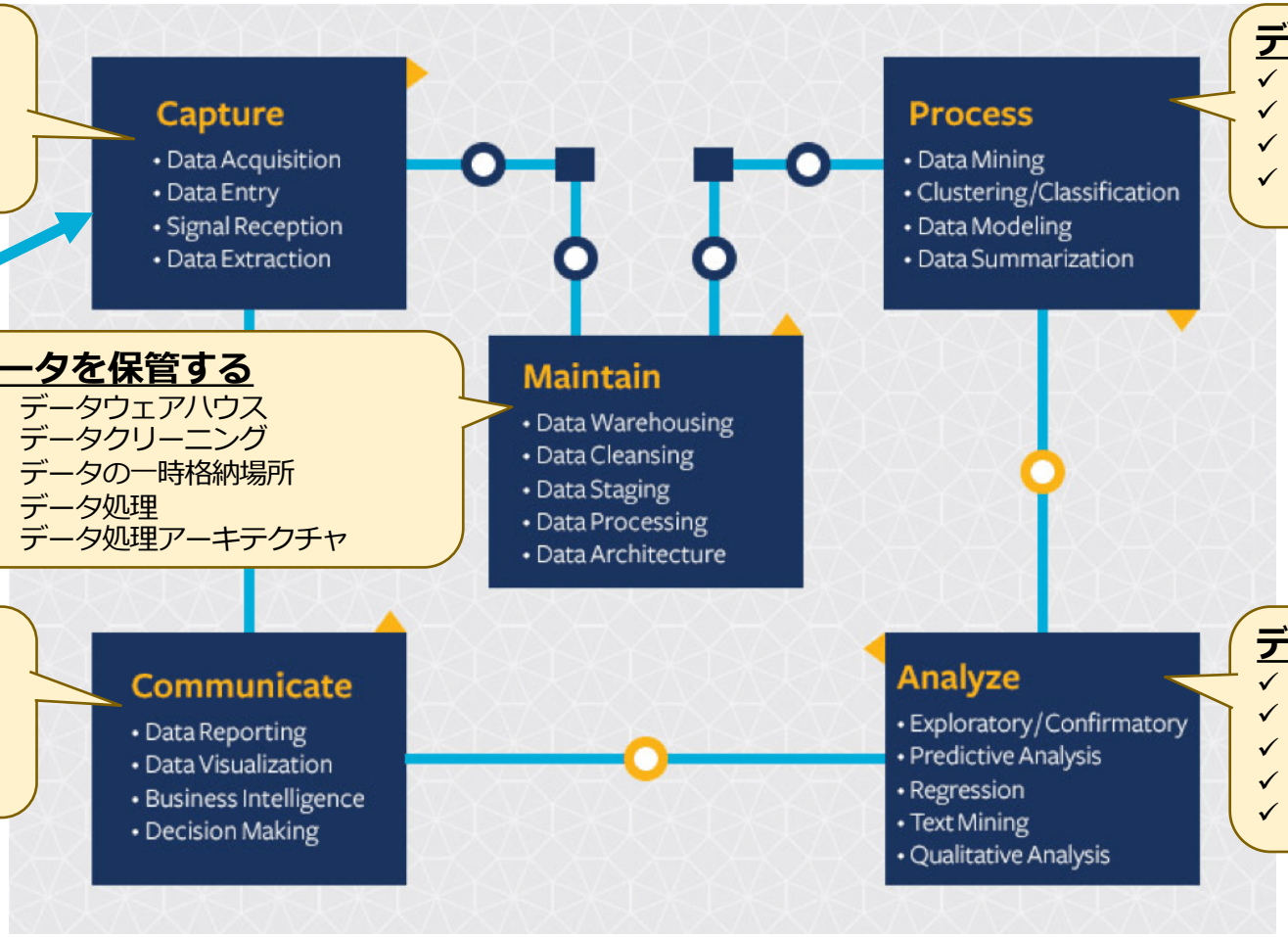
1. 自己紹介
2. データサイエンティストとは？
3. 機械学習とは？
4. データ分析コンテストの仕組み
5. 事業で活躍するKaggler
6. おわりに

データサイエンティストとは？

コンセンサスの取れた定義は現時点で世の中にはない。カリフォルニア大学バークレー校の情報学部によるData Science Life Cycleに沿って概観してみる。

目的設定
・何を知りたいか
・どのように使いたいか
など

データを手に入れる
✓ データを得る
✓ データを入力する
✓ シグナル受信
✓ データ抽出



データを処理する
✓ データマイニング
✓ クラスタリング/分類
✓ データモデリング
✓ データ集約、サマリー

データを保管する
✓ データウェアハウス
✓ データクリーニング
✓ データの一時格納場所
✓ データ処理
✓ データ処理アーキテクチャ

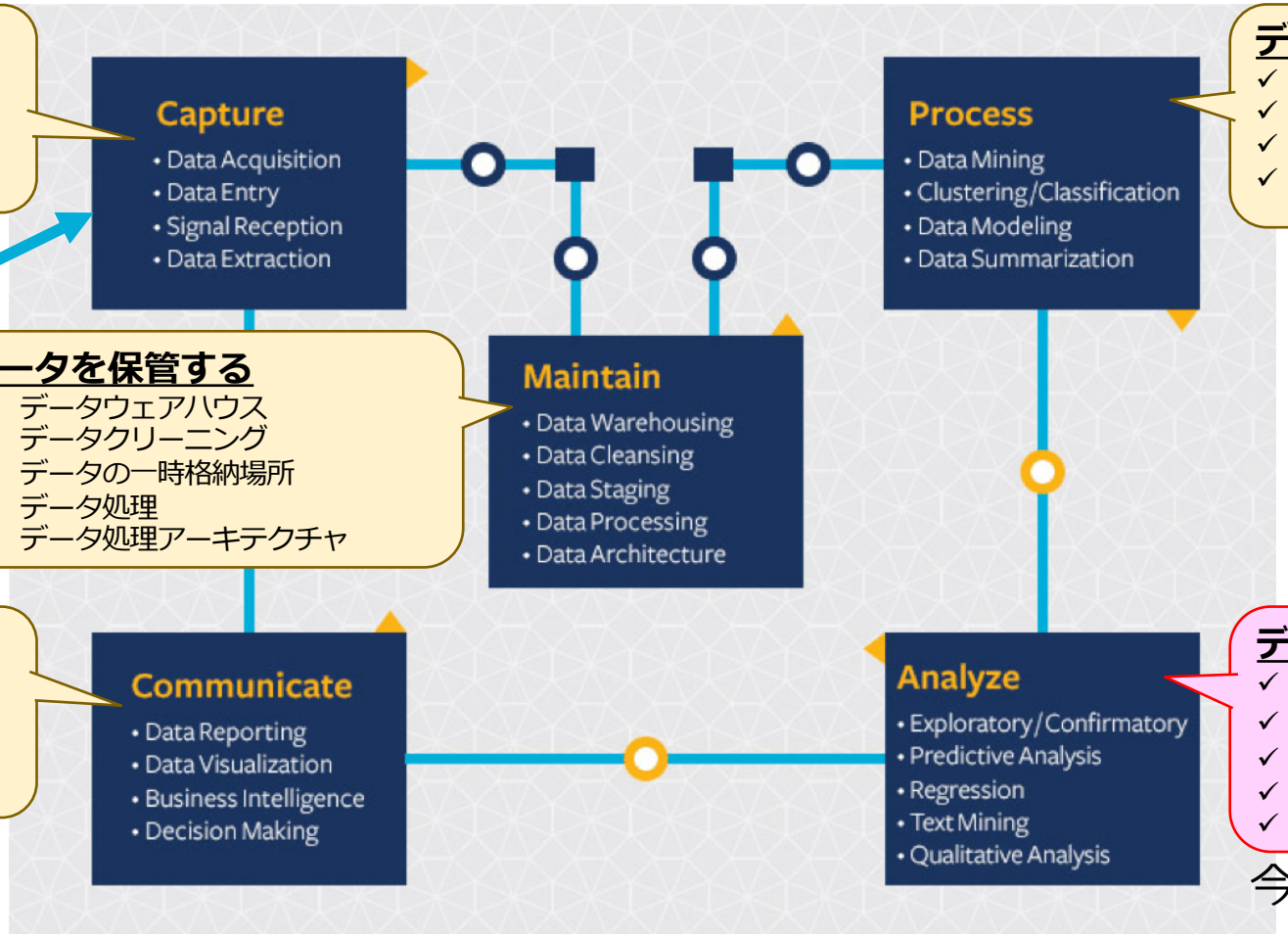
コミュニケーション
✓ レポーティング、報告
✓ データ可視化
✓ BIツール
✓ 意思決定

データを分析する
✓ 探索する/統計処理する
✓ 予測分析する
✓ 回帰分析
✓ テキストマイニング
✓ 質的分析

出典: https://datascience.berkeley.edu/about/what-is-data-science/

データサイエンティストとは？

コンセンサスの取れた定義は現時点で世の中にはない。カリフォルニア大学バークレー校の情報学部によるData Science Life Cycleに沿って概観してみる。



データを手に入れる
- ✓ データを得る
- ✓ データを入力する
- ✓ シグナル受信
- ✓ データ抽出

目的設定
- ・何を知りたいか
- ・どのように使いたいか

など

データを処理する
- ✓ データマイニング
- ✓ クラスタリング/分類
- ✓ データモデリング
- ✓ データ集約、サマリー

データを保管する
- ✓ データウェアハウス
- ✓ データクリーニング
- ✓ データの一時格納場所
- ✓ データ処理
- ✓ データ処理アーキテクチャ

コミュニケーション
- ✓ レポーティング、報告
- ✓ データ可視化
- ✓ BIツール
- ✓ 意思決定

データを分析する
- ✓ 探索する/統計処理する
- ✓ 予測分析する ⇒ **機械学習**
- ✓ 回帰分析
- ✓ テキストマイニング
- ✓ 質的分析

今日はここを重点的に

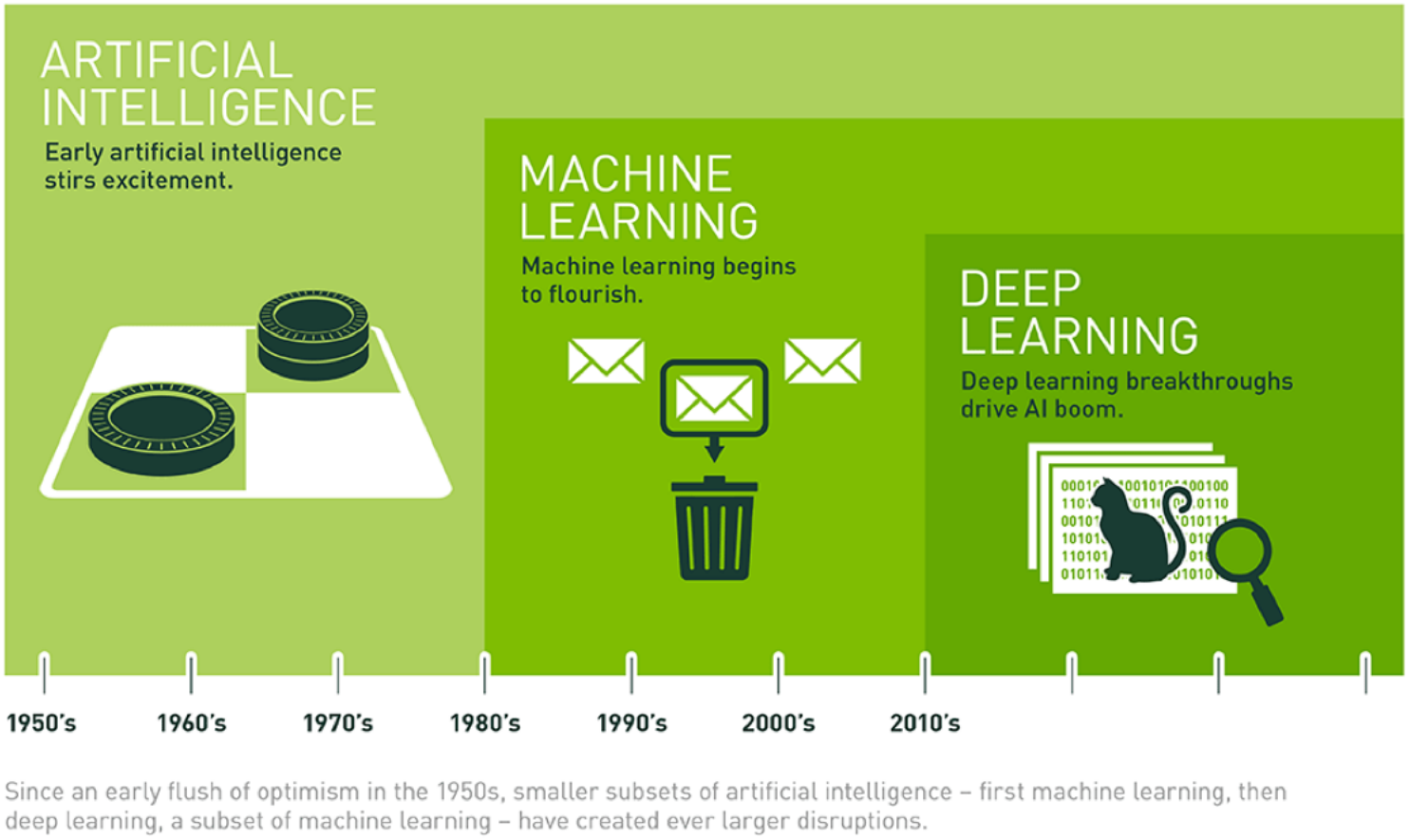
出典: https://datascience.berkeley.edu/about/what-is-data-science/

アジェンダ

1. 自己紹介
2. データサイエンティストとは？
3. 機械学習とは？
4. データ分析コンテストの仕組み
5. 事業で活躍するKaggler
6. おわりに

人工知能、機械学習、深層学習

昨今、人工知能、機械学習(Machine Learning)、深層学習(Deep Learning)という言葉が流行っているが、機械学習は人工知能の一分野。機械学習は「統計的学習」と呼ばれたりもする。



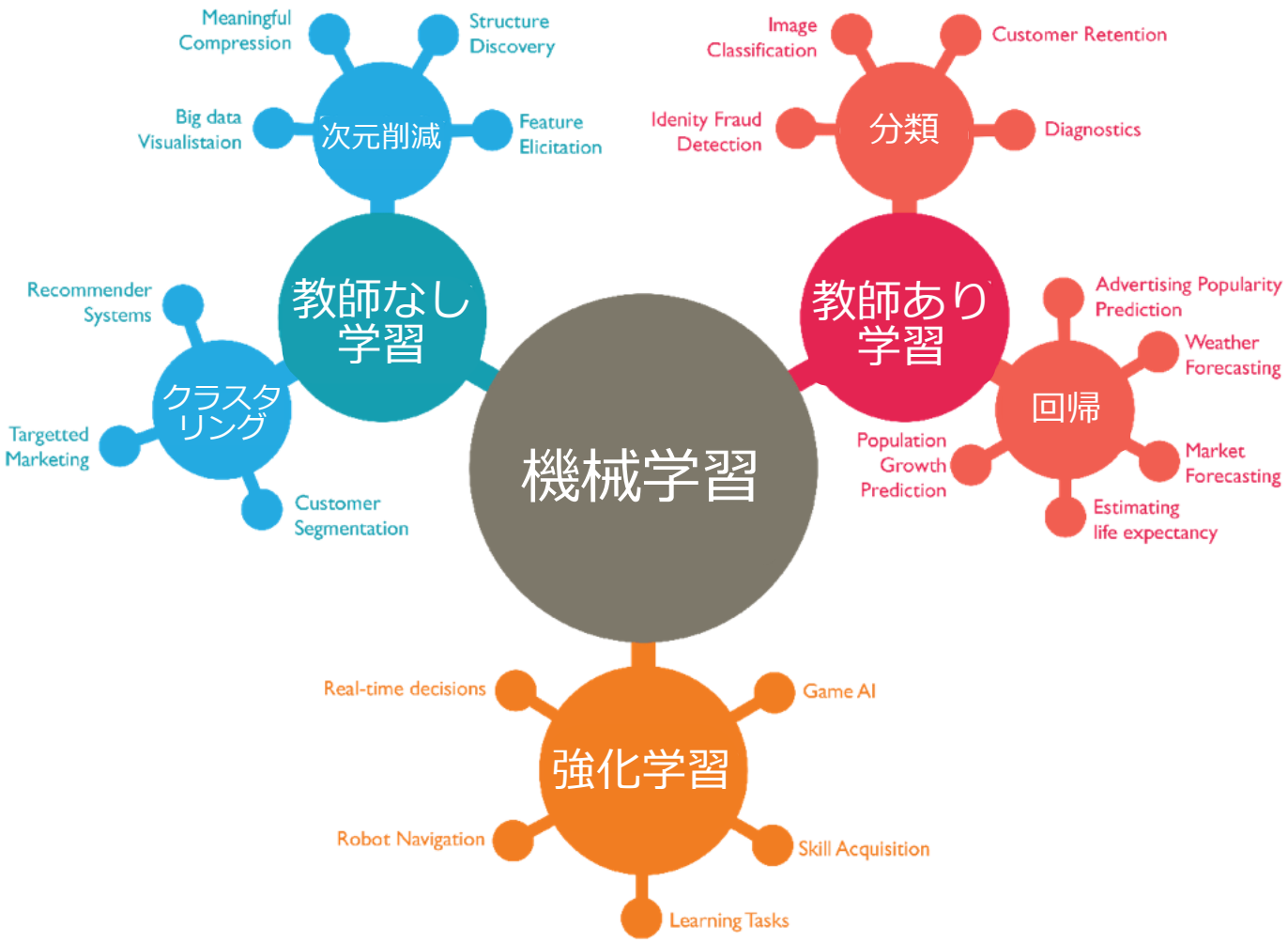
人は新たな知識や経験から学習を行う。コンピュータはデータをインプットして学習を行う。学習は大量なデータから法則を見つけ出すこととも言える。

出典: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

※ 人工知能とは？については人工知能学会の記事「教養知識としてのAI 〔第1回〕AIってなに？」も参考。
https://www.ai-gakkai.or.jp/comic_no1/

機械学習の分類

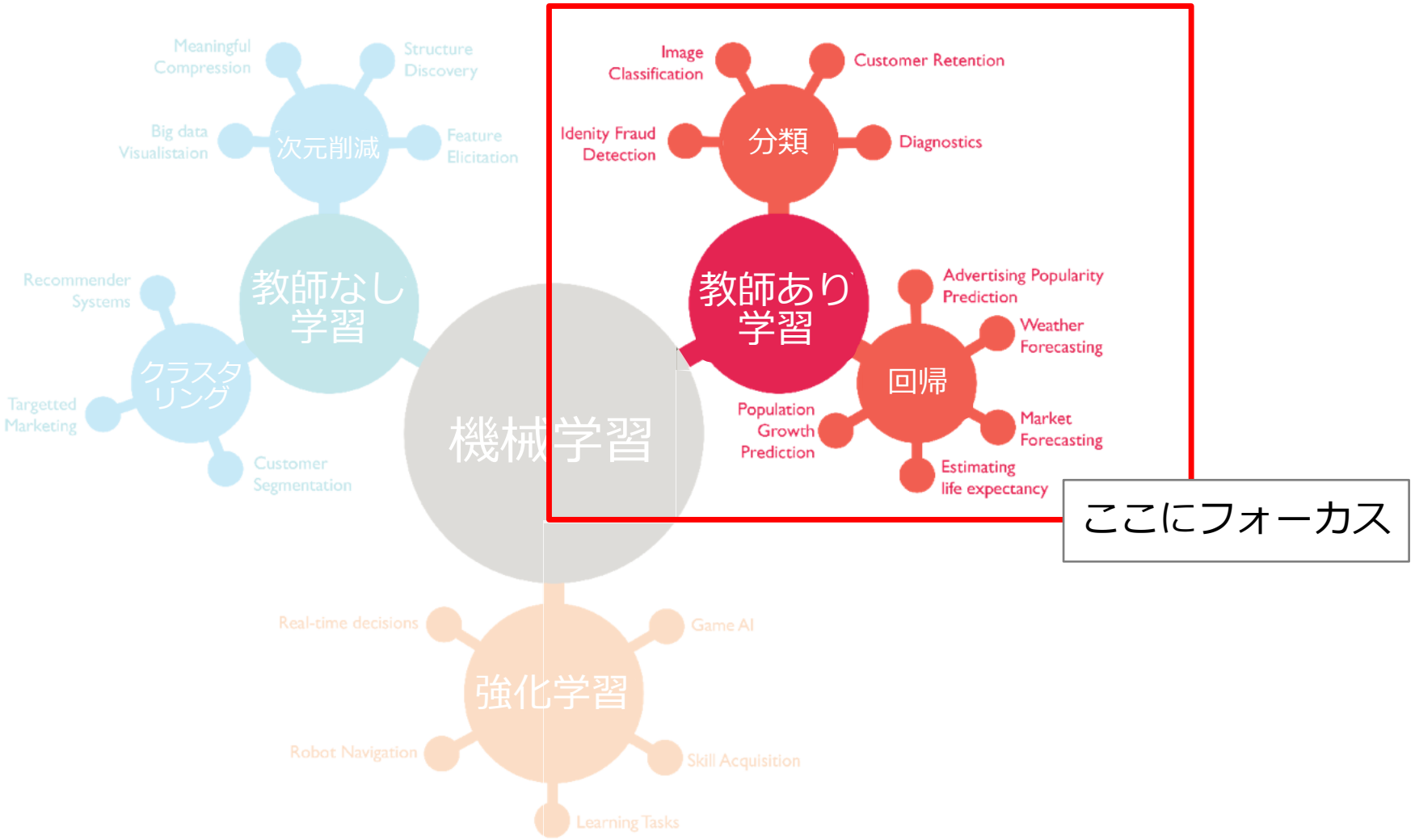
今日の講演ではデータサイエンティストのコアスキルの１つ、機械学習にフォーカスする。機械学習は下記のように分類できる。



出典: https://mse238blog.stanford.edu/2017/07/choftun/the-building-blocks-of-machine-learning/ を元に一部日本語化

機械学習の分類

今日の講演ではデータサイエンティストのコアスキルの１つ、機械学習にフォーカスする。機械学習は下記のように分類できる。



出典: https://mse238blog.stanford.edu/2017/07/choftun/the-building-blocks-of-machine-learning/ を元に一部日本語化

教師あり学習に利用するデータの例

Lending Club Loan Dataを例にとる。Excelのシートで表現できるようなデータ(テーブルデータ)。
機械学習に応用する際はここからインプットとなる学習データと、予測対象のデータを選ぶ。

学習データの例 （このようなデータが200万行）

loan_amnt ローン額	annual_inc 年収	int_rate 利子	term 返済期間	emp_title 職種	emp_length 勤続年数	home_ownership 住居所有種別	addr_state 居住州	last_pymnt_d 最終返済日	loan_condition ローン状況
30000	100,000.00	22.35	36 months	Supervisor	5 years	MORTGAGE	CA	Jan-19	Good Loan
40000	45,000.00	16.14	60 months	Assistant to the Treasurer (Payroll)	< 1 year	MORTGAGE	OH	Feb-19	Good Loan
8000	55,000.00	6.46	36 months	Meat Cutter	10+ years	MORTGAGE	WA	NaN	Bad Loan

インプット

答え

予測対象データの例

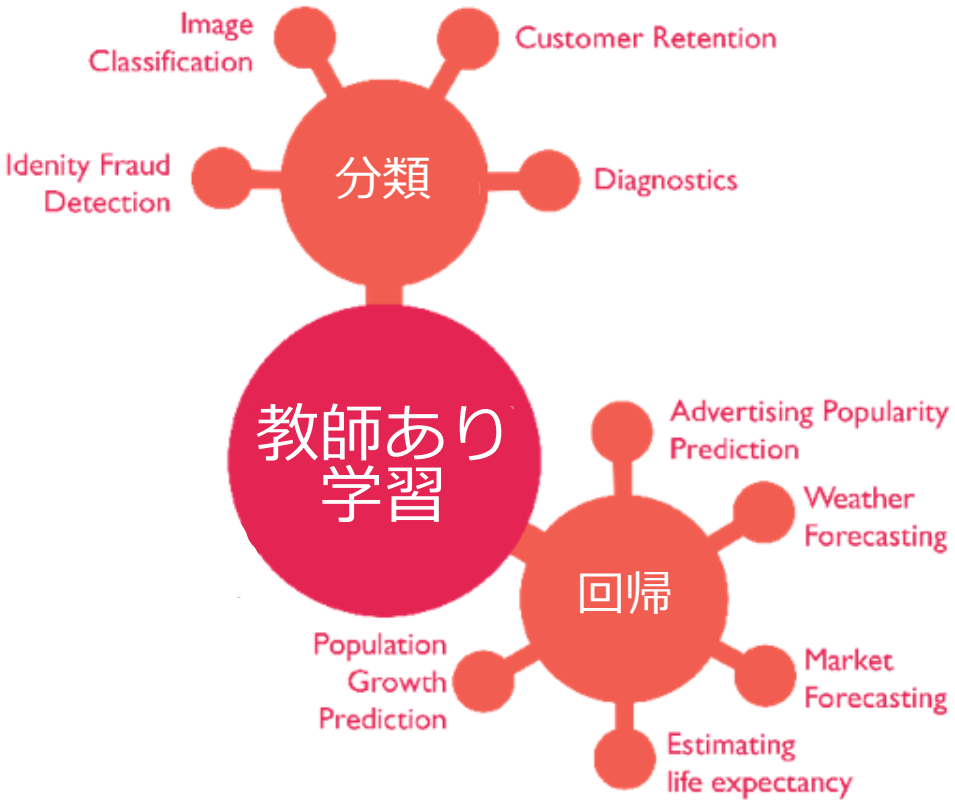
6000	17,000.00	14.47	36 months	NaN	NaN	MORTGAGE	FL	NaN	?
7500	50,000.00	12.73	36 months	NaN	NaN	MORTGAGE	IN	Oct-18	?
30000	109,000.00	20.89	36 months	President	8 years	MORTGAGE	TX	Feb-19	?

まだ答えがわかっていない
データを予測したい

出典： https://www.kaggle.com/kenmatsu4/lending-club-data-analysis

教師あり学習

教師あり学習とは、人が事前にインプットデータと答えのペアを用意しておき、このペアの関係性を**機械学習モデル**に学習させるもの。答えの種類に応じて分類と回帰の２種類に分けられる。

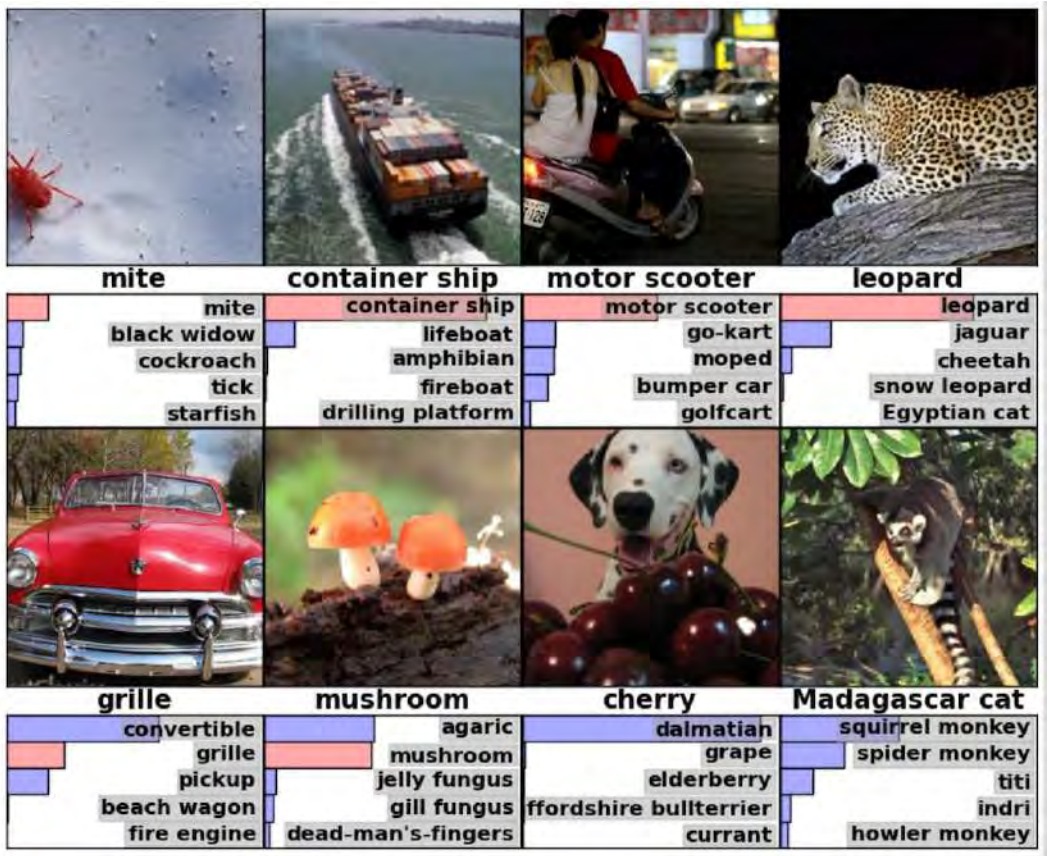


	内容	例
分類	答えが有限個の種類に分けられ、そのどれに該当するかを予測する問題	✓ ローンの申し込み情報から、顧客が返済可能かどうかを当てる問題（二値分類） ✓ 画像認識（画像分類 ex 犬か猫か、うさぎか） 
回帰	予測値が数値で表される問題	✓ 気温、季節、イベント内容、などから来場者数を当てる問題 ✓ 不動産の情報から価格を予測する問題

画像分類タスクにおける機械学習

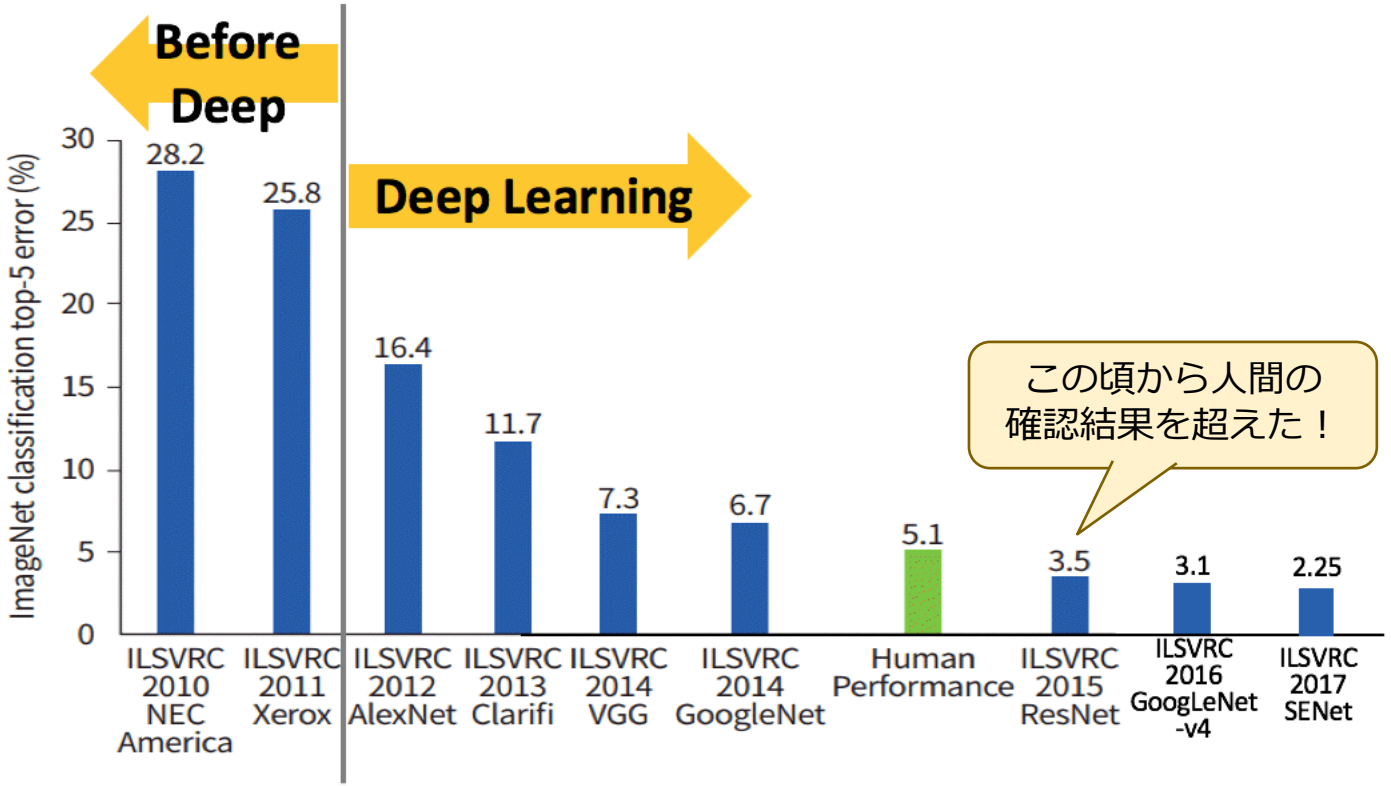
画像分類の分野ではDeep Learningという技術の出現で予測精度が急激に向上し、人間の結果を超えるまでに。

画像分類タスク top 5 prediction



5つの予想結果に答えが入っていれば正解

Deep Learningの出現と分類精度向上の歴史



https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

https://www.pfmjournal.org/m/journal/view.php?number=32

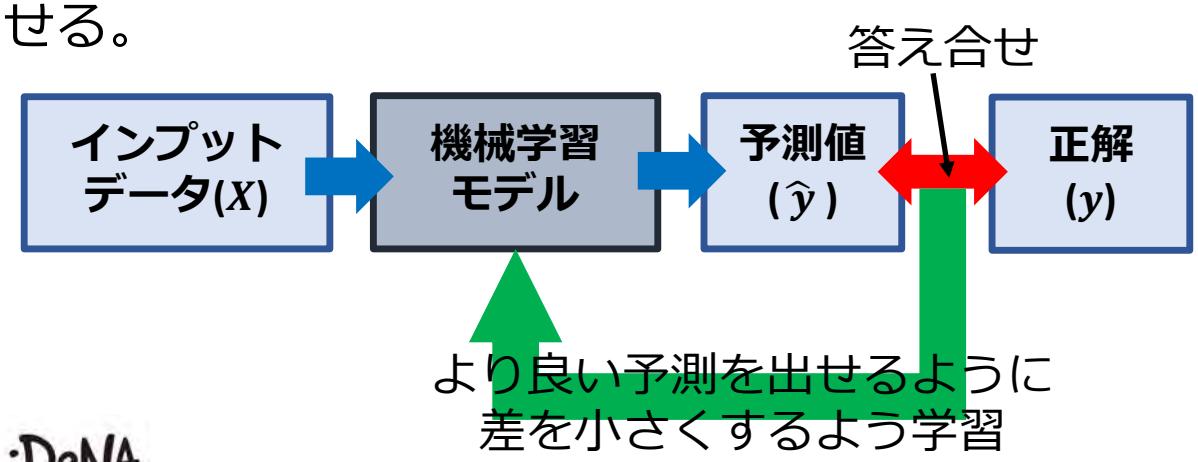
機械学習モデルの活用

下記のステップで機械学習のモデルの学習を行って活用する。インプットデータを得た時点ではすぐに正解データが得られないような問題設定に活用できる。

機械学習活用のステップ

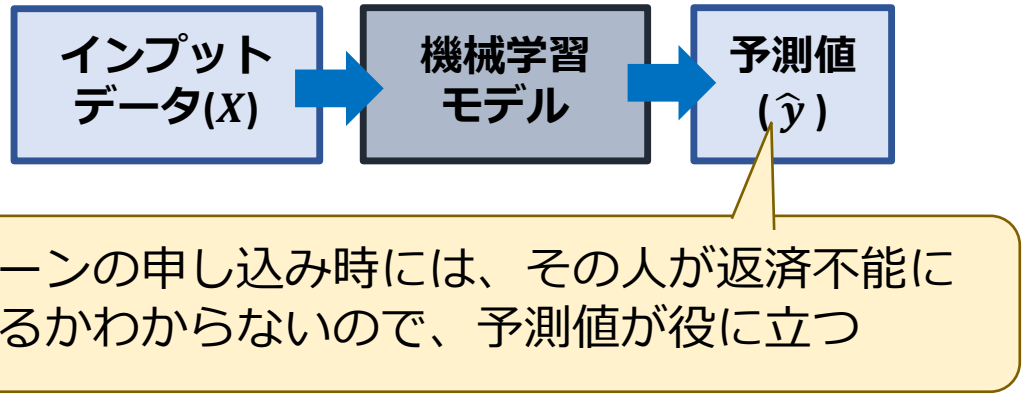
学習・評価

予測と正解が合うように機械学習モデルを学習させる。



活用(推論)

構築済みの機械学習モデルを用いてインプットデータから予測値を算出、活用する。



DeNA

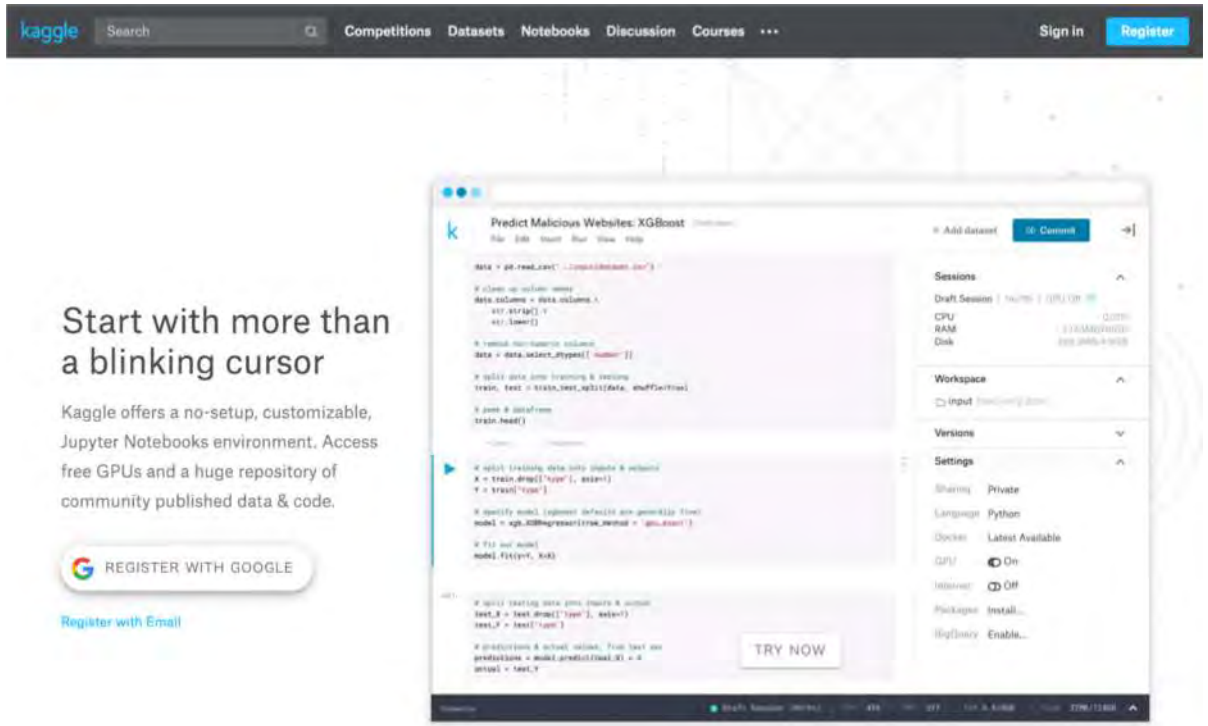
アジェンダ

1. 自己紹介
2. データサイエンティストとは？
3. 機械学習とは？
4. データ分析コンテストの仕組み
5. 事業で活躍するKaggler
6. おわりに

データ分析コンテスト

広く知られているデータ分析コンテストには下記の２つがある。データマイニング学会 KDDが主催するKDD CUPも有名。最近は個別の企業がコンテストを開催することも。

Kaggle



世界最大のデータ分析コンテスト。約350万人が登録し、12万人以上が実際にコンテストに参加実績がある。世界中からデータサイエンティストが集う。

https://www.kaggle.com/

SIGNATE

日本のデータ分析コンテスト。日本語で提供されており、国や日本企業からの出題も。

https://signate.jp/

DeNA Kaggle社内ランク制度

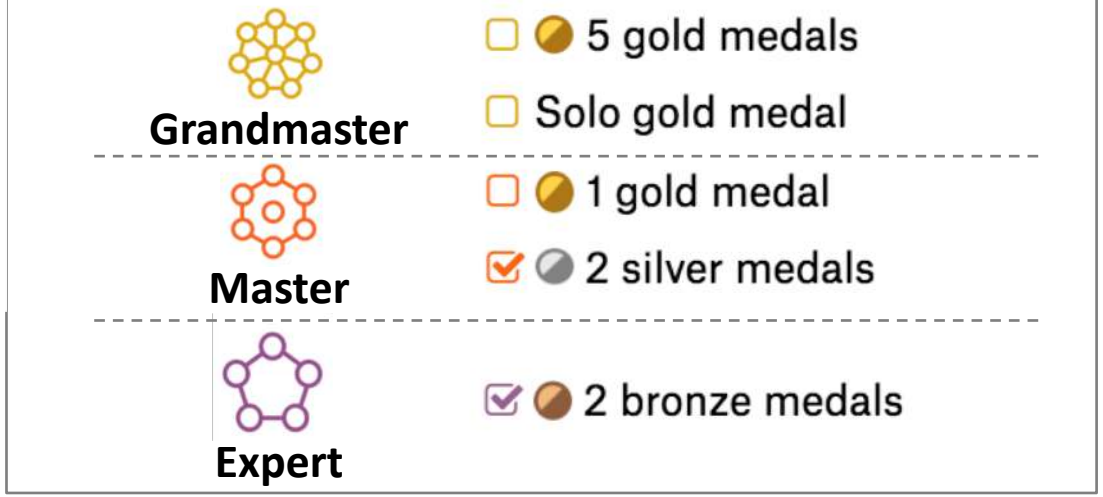
Kaggleの成績に応じて業務時間の一定割合をKaggleに当てることが可能

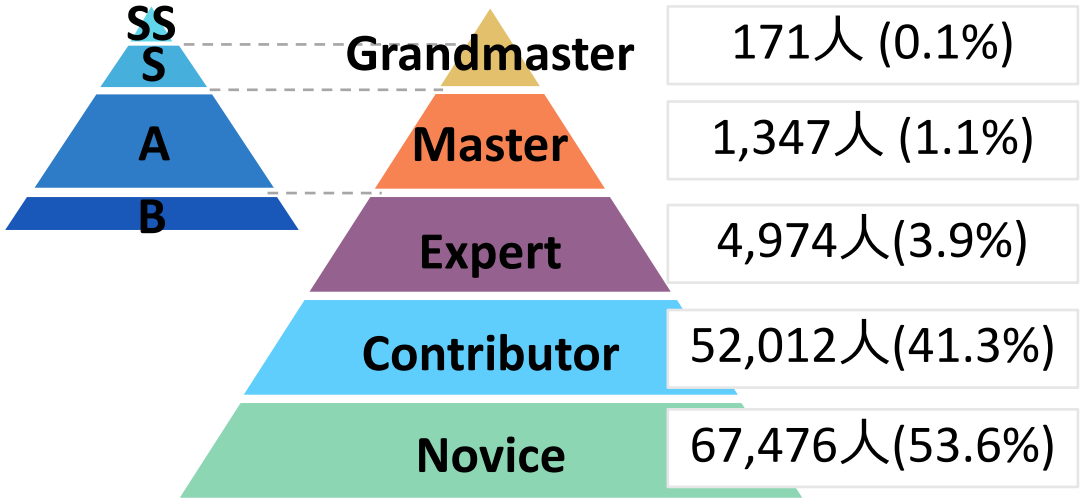
Competition Medals

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

https://dena.ai/kaggle/

Performance Tiers





competition point >0 の人数 (2019年12月現在)

https://www.kaggle.com/progression

【参考】Kaggler の最近の実績（抜粋）

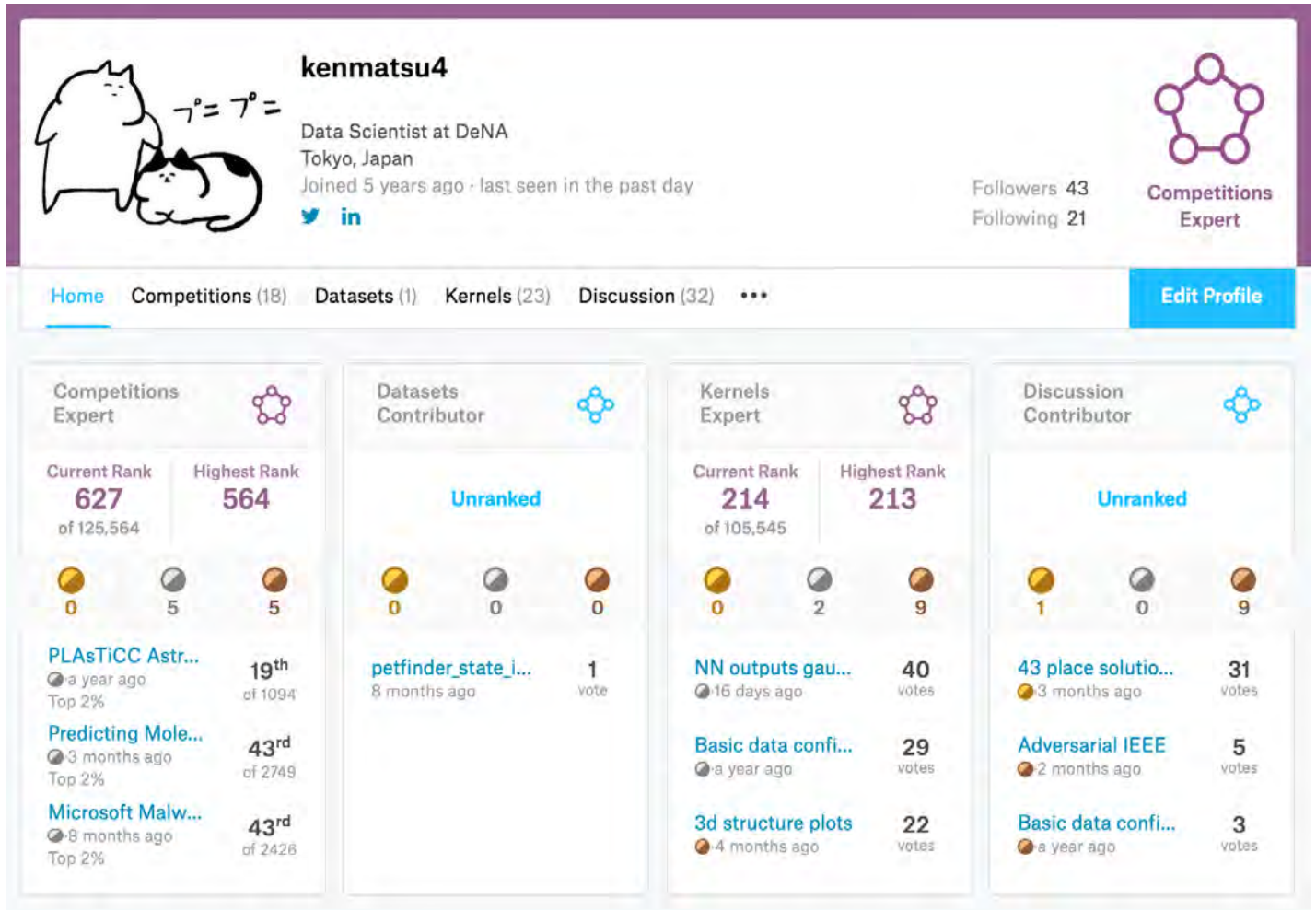
DeNA x AI News
https://dena.ai/news/

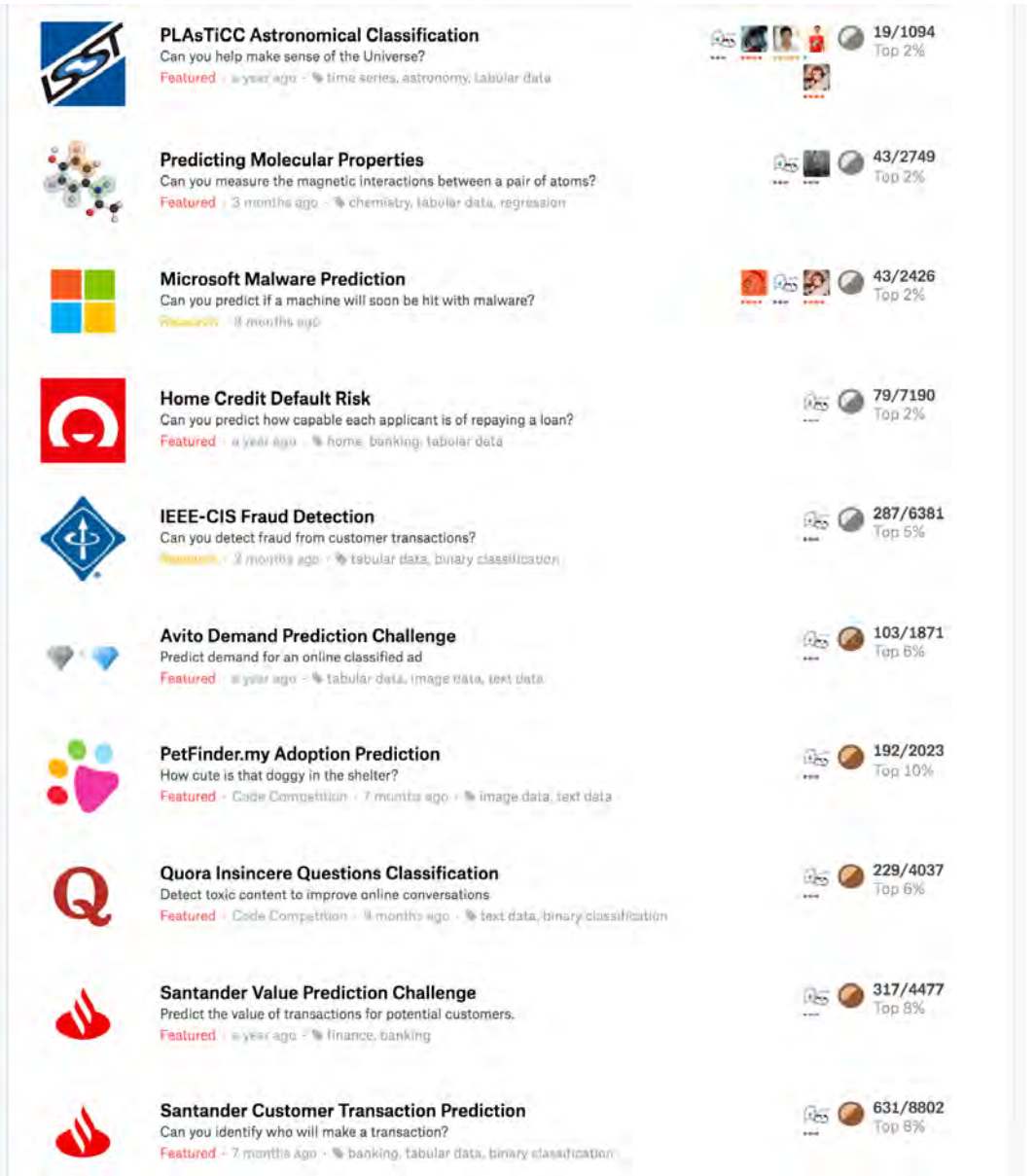
様々なコンペで上位入賞

2019.11.15	Kaggleコンペティション”RSNA Intracranial Hemorrhage Detection”でDeNAの大越がソロ3位入賞 & Kaggle Grandmasterになりました
2019.10.25	北京で開催された「Kaggle Days China」のオフラインコンペティションで、DeNAのデータサイエンティストを含むチームが優勝しました
2019.10.8	Kaggleコンペティション「IEEE-CIS Fraud Detection」で、DeNAのデータサイエンティストを含むチームが8位に入りました
2019.10.7	Open Images 2019 コンペティション 物体検出部門で、DeNAのAI研究開発エンジニアが6位（ソロ参加者中1位）を獲得しました
2019.9.30	Kaggleコンペティション "Recursion Cellular Image Classification" にてDeNAのAI研究開発エンジニアを含むチームが4位に入りました
2019.4.19	Kaggleコンペティション「Santander Customer Transaction Prediction」で、DeNAのデータサイエンティストを含む3チームが2位、8位、9位に入りました

分析コンテスト戦歴

Kaggleで銀メダル５つ（ソロ２つ、チーム３つ） 2019年11月現在 125,564人中 627位 (top 0.5%)





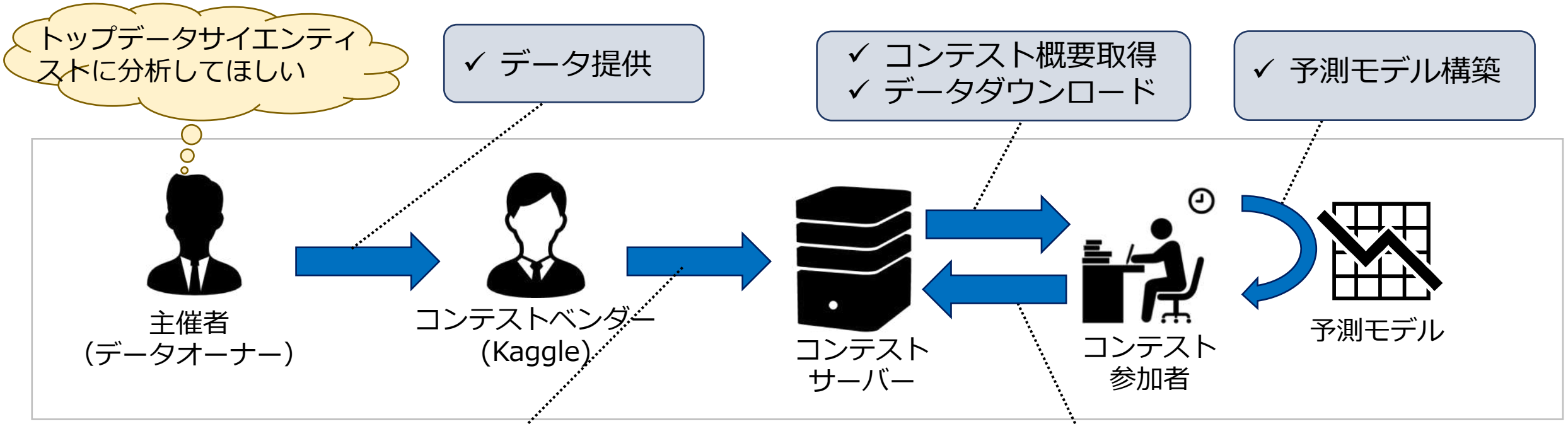
日本での実績 SIGNATE

✓ 産業技術総合研究所 衛星画像分析コンテスト 2位入賞

https://www.slideshare.net/matsukenbook/signate-108228406

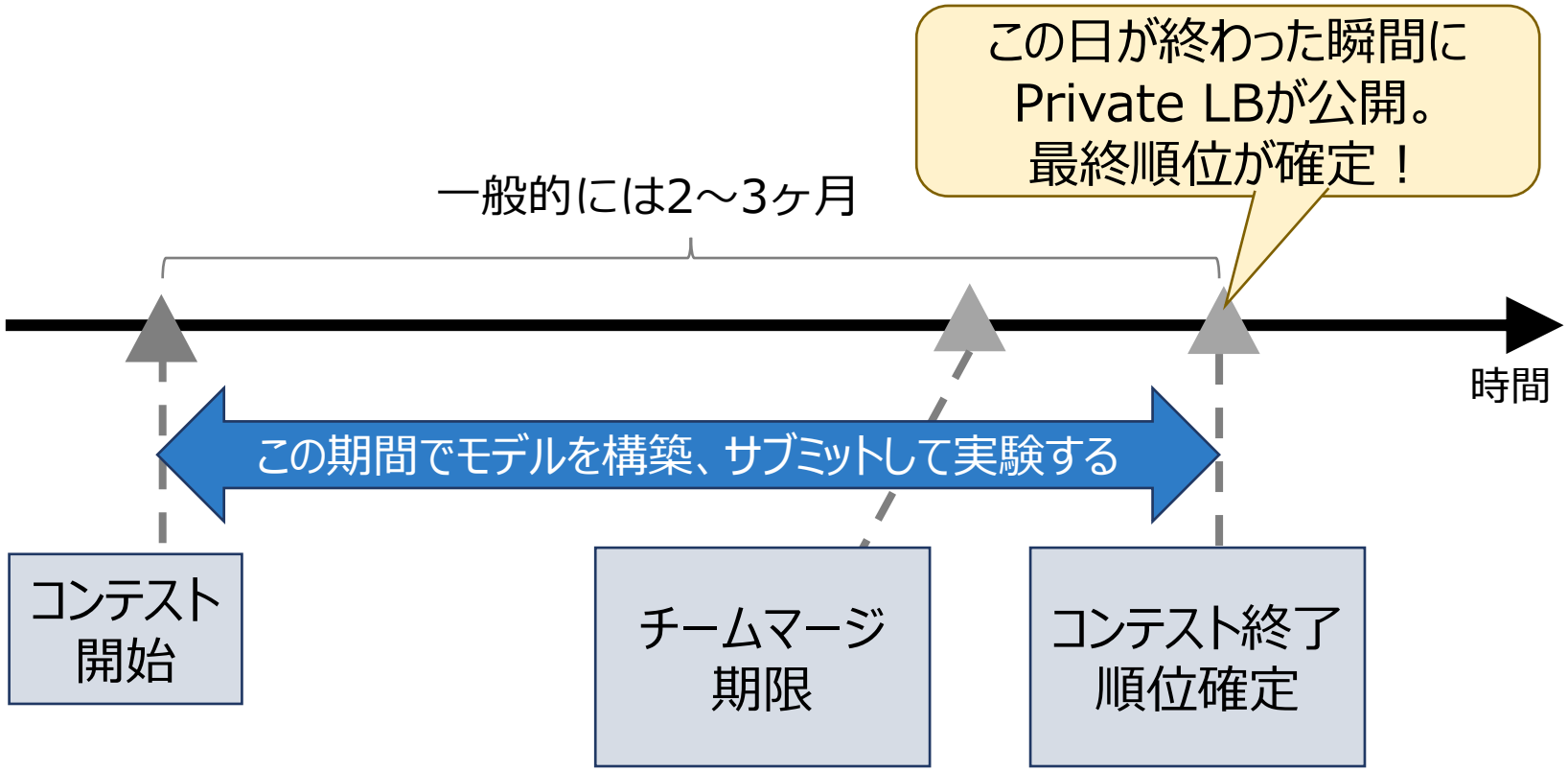
データ分析コンテストとは

主催者より分析対象データが提供され、コンテストベンダーが運営。参加者は自由に参加でき、作成した予測モデルの予測精度を競う。決められた期日に最終精度スコアに基づき、順位・入賞者を決定。



コンテストのタイムスケジュール

典型的なKaggleのコンテストは2～3ヶ月の開催期間。コンテスト最終日の１週間くらい前に新規参加の停止と、チームマージ（複数人でチームを組むこと）の期限があり、最終日を過ぎた直後に最終順位が確定する



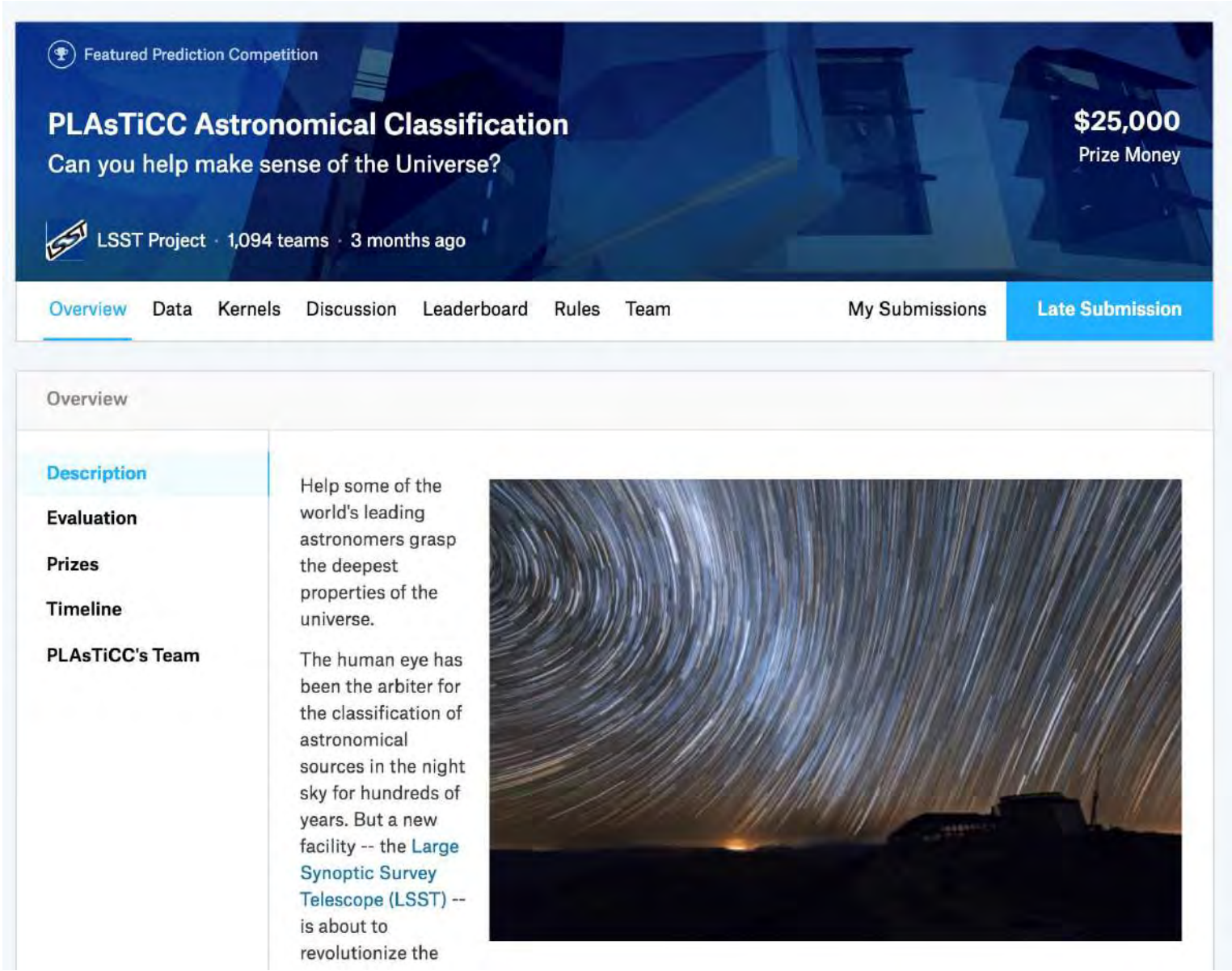
Private Leader Board
@コンテストサーバー

順位	ユーザ名	スコア	応募件数	投稿日時
1	user_001	0.834	48	2018/4/26 20:07
2	Kenmatsu4	0.812	95	2018/4/26 20:40
3	user_002	0.807	2	2018/4/23 14:37
4	user_003	0.798	84	2018/4/26 20:05
5	user_004	0.791	40	2018/4/24 22:58
6	user_005	0.791	6	2018/4/26 21:22
7	user_006	0.790	14	2018/4/25 21:03
⋮	⋮	⋮	⋮	⋮

リーダーボードに全参加者の
精度スコアがランキングで表示される

✓開催期間中、リーダーボードには予測精度スコアが記載されランキング表示される。
✓1日あたりのサブミット回数は数回に制限されている

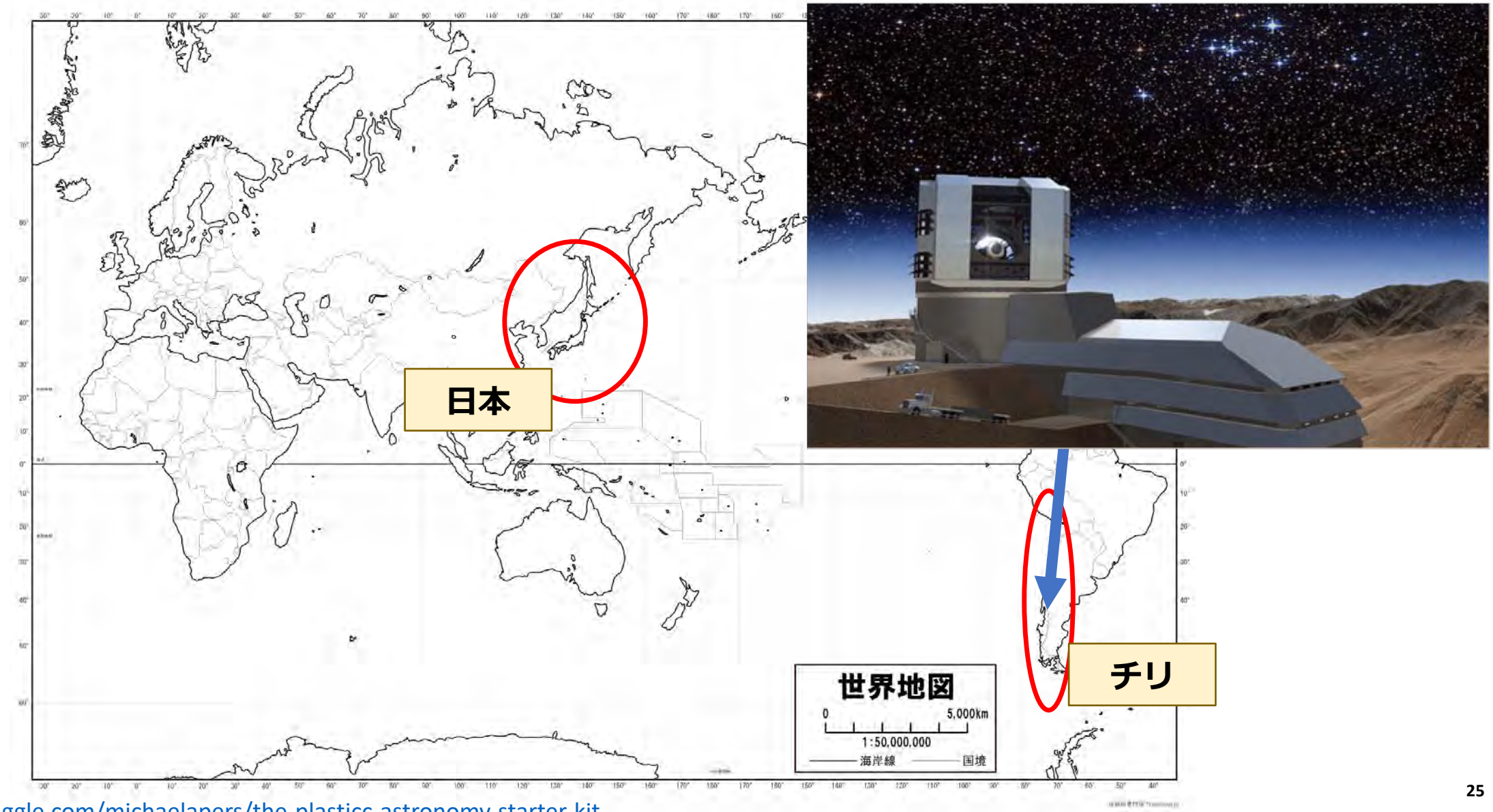
コンテスト例：天体観測で見つかった物体の識別



https://www.kaggle.com/c/PLAsTiCC-2018

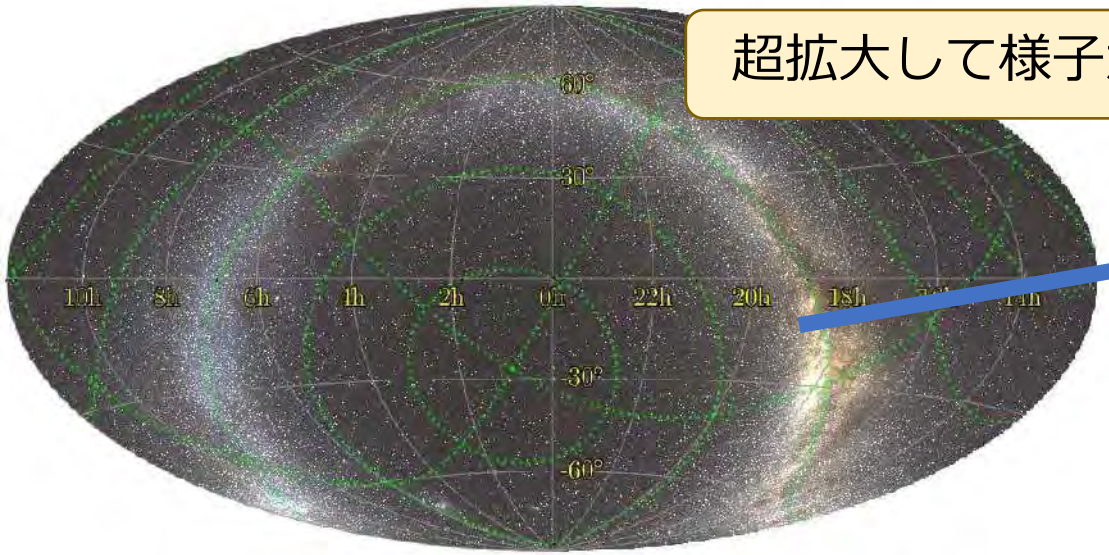
超高性能天体望遠鏡 LSST

2019 年に完成する超高性能天体望遠鏡の観測天体を 15 種類に分類



https://www.kaggle.com/michaelapers/the-plasticc-astronomy-starter-kit

広大な宇宙を超拡大して様子を観察できる



超拡大して様子がわかる

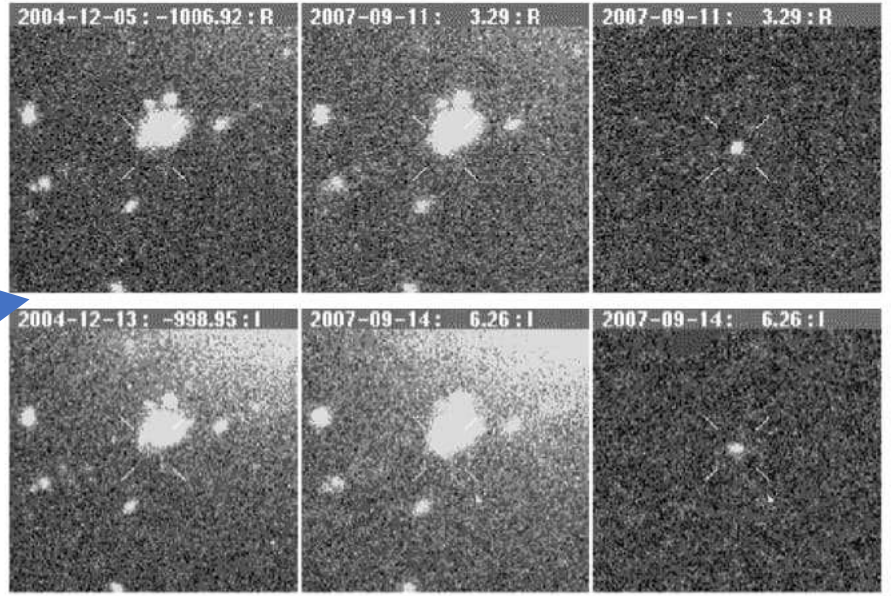


Figure 14: Difference Imaging "Postage Stamps" (Reference, Image, Difference)

14種類+その他 計15種の分類

TABLE 1
SUMMARY OF TRANSIENT AND VARIABLE MODELS FOR PLAsTiCC.

Model Class Num[a]: Name	Model Description	Contributor(s)[b]	N_{event} Gen[c]	N_{event} Train[d]	N_{event} Test[e]	Redshift Range[f]
90: SNIa	WD detonation, Type Ia SN	RK	16,353,270	2,313	1,659,831	< 1.6
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	< 0.9
52: SNIax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	< 1.3
42: SNII	Core collapse, Type II SN	SG,LG:RK,JRP:VAV	59,198,660	1,193	1,000,150	< 2.0
62: SNIbc	Core collapse, Type Ibc SN	VAV:RK,JRP	22,599,840	484	175,094	< 1.3
95: SLSN-I	Super-lum. SN (magnetar)	VAV	90,640	175	35,782	< 3.4
15: TDE	Tidal disruption event	VAV	58,550	495	13,555	< 2.6
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	< 0.3
88: AGN	Active galactic nuclei	SD	175,500	370	101,424	< 3.4
92: RRL	RR Lyrae	SD	200,200	239	197,155	0
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	0
16: EB	Eclipsing binary stars	AP	220,200	924	96,572	0
53: Mira	Pulsating variable stars	RH	1,490	30	1,453	0
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2,820	151	1,303	0
991: μLens-Binary	μ-lens from binary lens	RD,AA	1,010	0	533	0
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1,702	< 0.4
993: CaRT	Calcium-rich Transient	VAV	2,834,500	0	9,680	< 0.9
994: PISN	Pair-instability SN	VAV	5,650	0	1,172	< 1.9
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	0
TOTAL	Sum of all models		117,128,700	7,846	3,492,888	—

class53: 脈動変光星

class90
Ia型超新星

https://www.kaggle.com/michaelapers/the-plasticc-astronomy-starter-kit

https://arxiv.org/pdf/1903.11756.pdf

https://en.wikipedia.org/wiki/Type_Ia_supernova

https://www.quantamagazine.org/variable-stars-have-strange-nonchaotic-attractors-20150310

データの概要

- ✓ 右記のようなデータが学習用データとして約140万行与えられる、天体数は7848天体
- ✓ 予測対象のデータはもっと多く約4.5億行。天体数は約350万天体
- ✓ この予測対象のデータの精度を競う
- ✓ その他、天体に関する属性情報と、正解データ：天体の種別番号が与えられる

天体のID　時刻　光の周波数帯　光の強さ　光の測定誤差

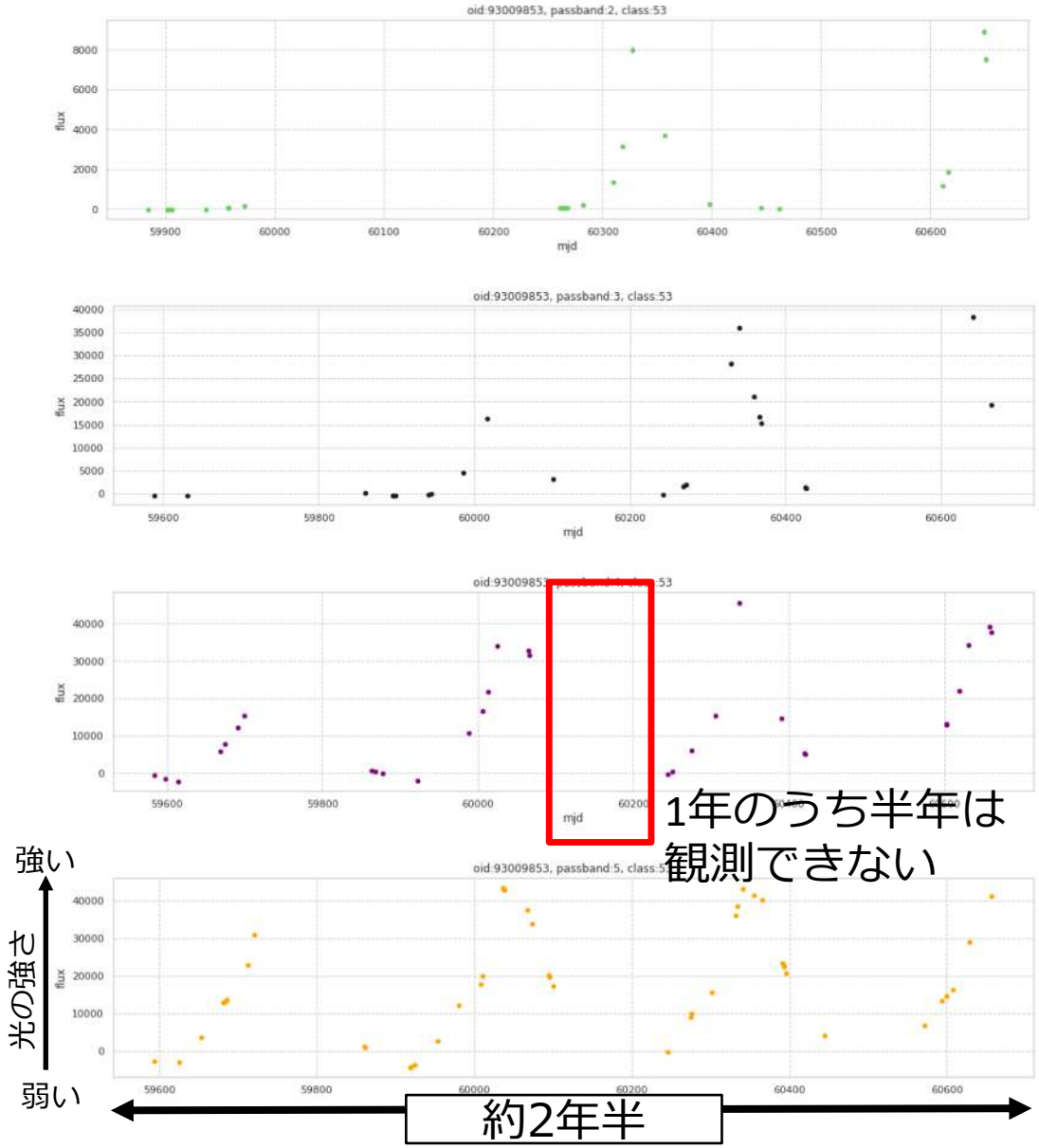
	object_id	mjd	passband	flux	flux_err	detected
0	615	59,750.42290	2	-544.81030	3.62295	1
1	615	59,750.43060	1	-816.43433	5.55337	1
2	615	59,750.43830	3	-471.38553	3.80121	1
3	615	59,750.44500	4	-388.98498	11.39503	1
4	615	59,752.40700	2	-681.85889	4.04120	1
5	615	59,752.41470	1	-1,061.45703	6.47299	1
6	615	59,752.42240	3	-524.95459	3.55275	1
7	615	59,752.43340	4	-393.48023	3.59935	1

⋮

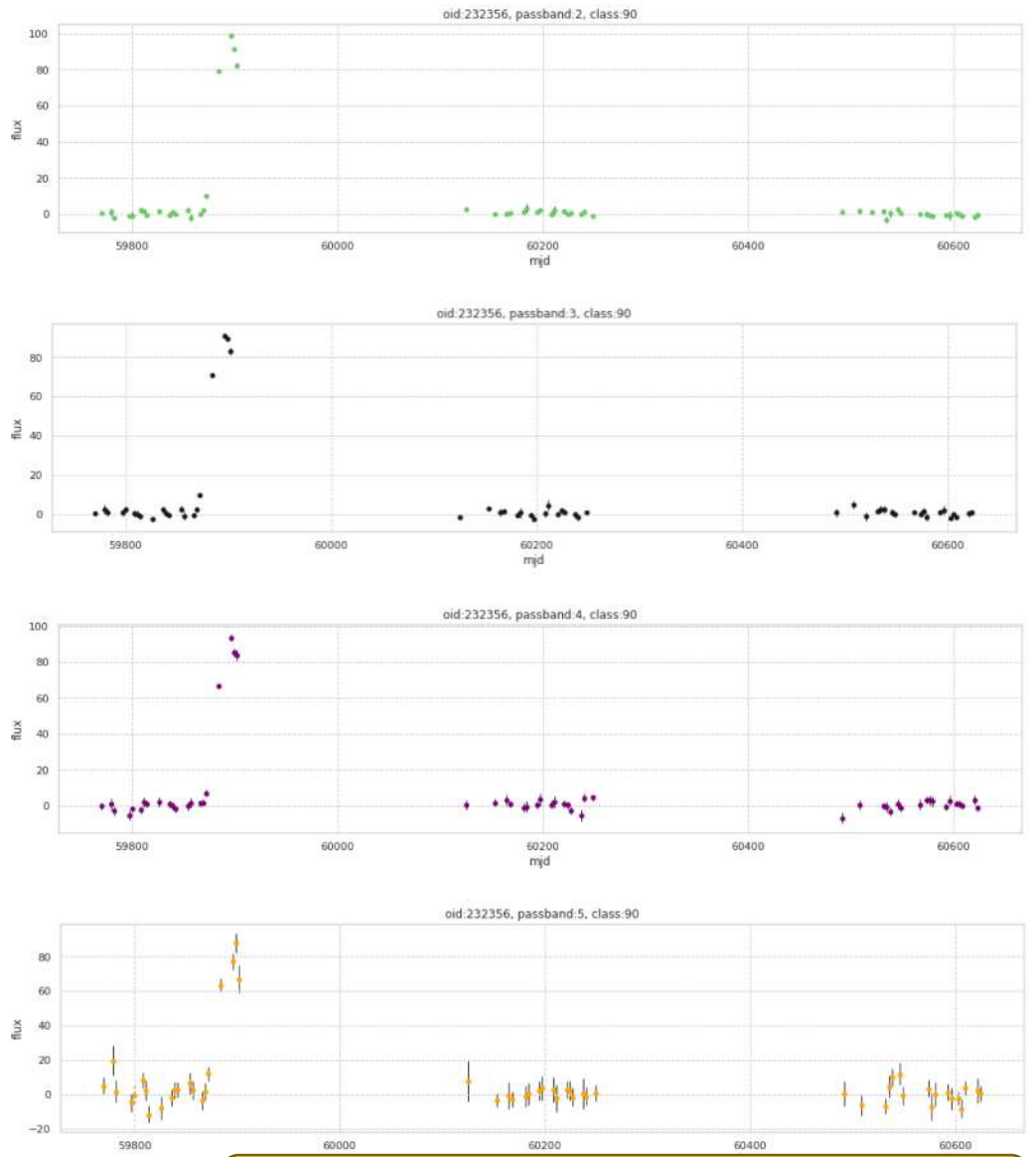
天体のID　正解データ：天体の種別番号

	object_id	ra	decl	gal_l	gal_b	ddf	hostgal_specz	hostgal_photoz	hostgal_photoz_err	distmod	mwebv	target	gal
0	615	349.04605	-61.94384	320.79653	-51.75371	1	0.00000	0.00000	0.00000	nan	0.01700	92	1
1	713	53.08594	-27.78440	223.52551	-54.46075	1	1.81810	1.62670	0.25520	45.40630	0.00700	88	0
2	730	33.57422	-6.57959	170.45559	-61.54822	1	0.23200	0.22620	0.01570	40.25610	0.02100	42	0
3	745	0.18987	-45.58666	328.25446	-68.96930	1	0.30370	0.28130	1.15230	40.79510	0.00700	90	0

class53: 脈動変光星



class90: Ia型超新星

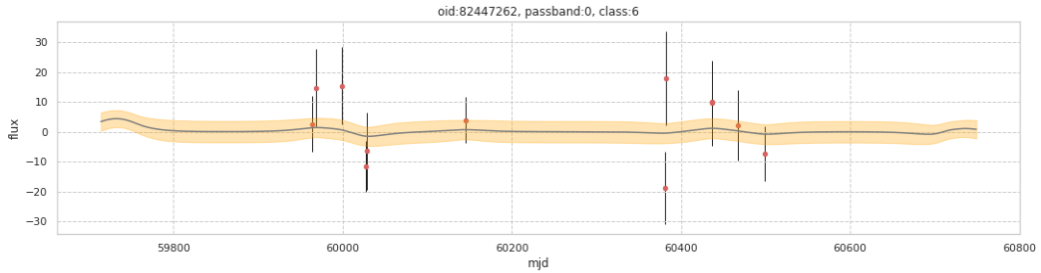


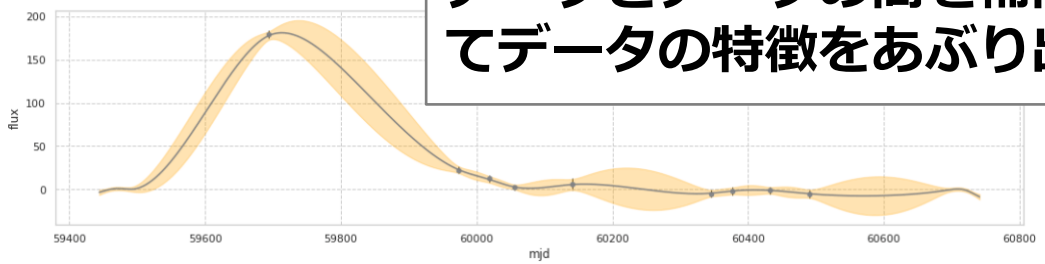
https://www.kaggle.com/kenmatsu4/feature-engineering-with-gaussian-process

Gaussian Process適用前

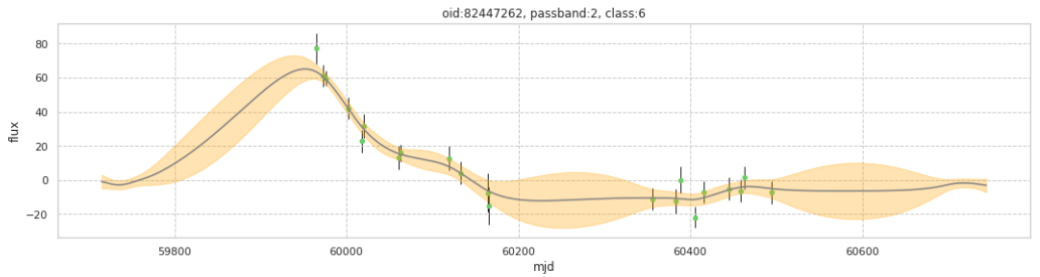
データ解析の例

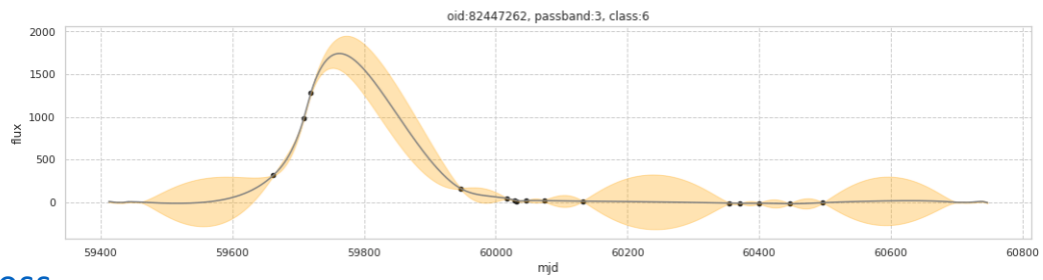
```
341
342     optimizer_list = ["L-BFGS-B",
343                       "Nelder-Mead",
344                       "Powell",
345                       "CG",
346                       "BFGS",
347                       "Newton-CG",
348                       "TNC",
349                       "COBYLA",
350                       "SLSQP",
351                       "dogleg",
352                       "trust-ncg",]
353
354     def neg_log_like(params, y, gp):
355         gp.set_parameter_vector(params)
356         return -gp.log_likelihood(y)
357
358     def grad_neg_log_like(params, y, gp):
359         gp.set_parameter_vector(params)
360         return -gp.grad_log_likelihood(y)[1]
361
362     def build_gp_model(x, y, yerr, n_param = 2):
363         log_sigma = 0
364         log_rho = 0
365         eps = 0.001
366         bounds = dict(log_sigma=(-15, 15), log_rho=(-15, 15))
367         kernel = terms.Matern32Term(log_sigma=log_sigma,
368                                     log_rho=log_rho,
369                                     eps=eps,
370                                     bounds=bounds)
371
372         gp = celerite.GP(kernel, mean=0)
373         gp.compute(x, yerr)
374
375         initial_params = gp.get_parameter_vector()
376         bounds = gp.get_parameter_bounds()
377
378         # depend on a combination of optimizer and dataset, minimize function throw exception,
379         # so trying all type of method.
380         for opt in optimizer_list:
381             try:
382                 r = minimize(neg_log_like,
383                              initial_params,
384                              jac=grad_neg_log_like,
385                              method=opt, #"L-BFGS-B",
386                              bounds=bounds,
387                              args=(y, gp))
388                 return gp
389
390             except Exception as e:
391                 pass
392         raise Exception("[build_gp_model] can`t optimize")
393
```





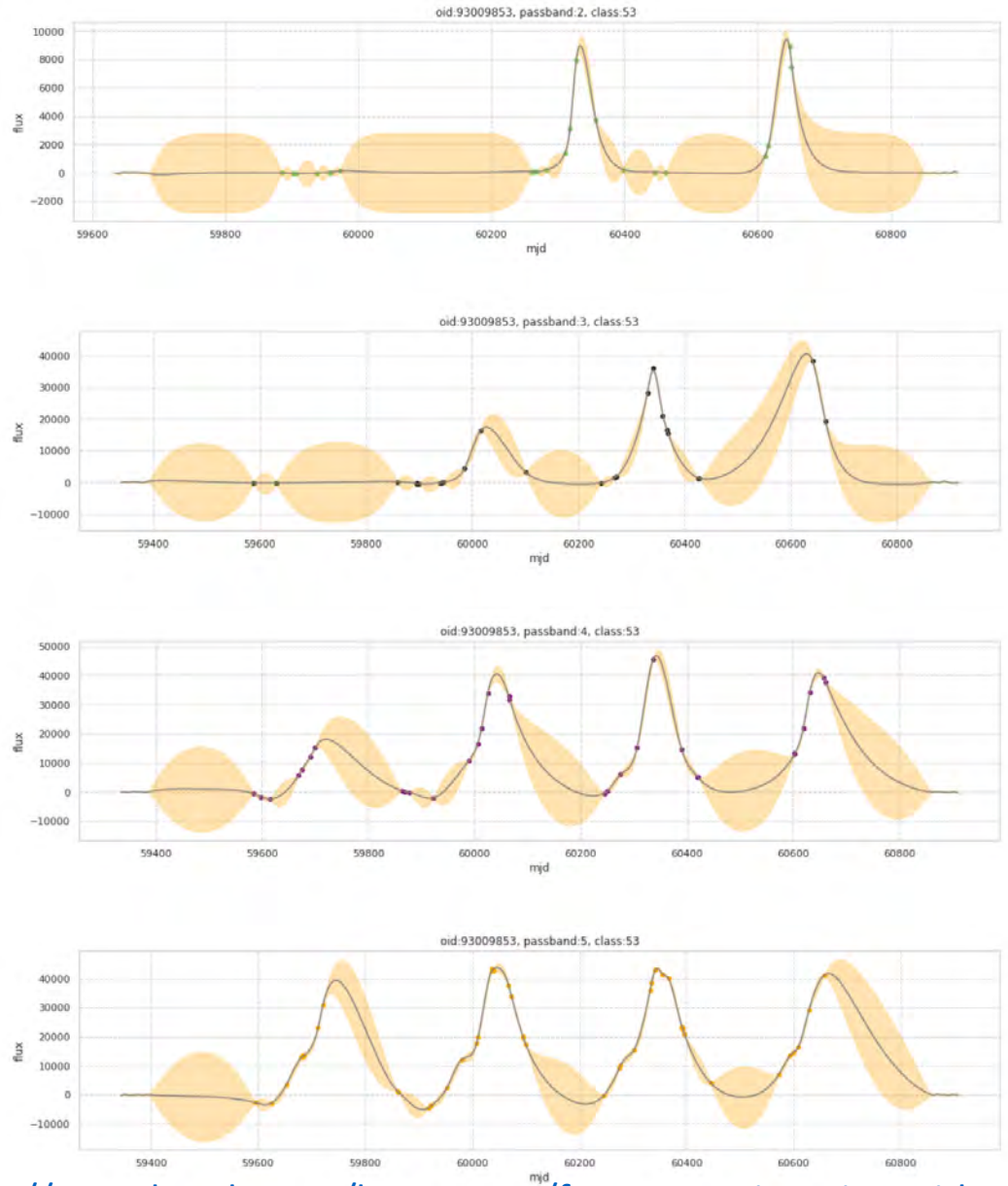
データとデータの間を補間してデータの特徴をあぶり出す





https://www.kaggle.com/kenmatsu4/feature-engineering-with-gaussian-process

class53: 脈動変光星



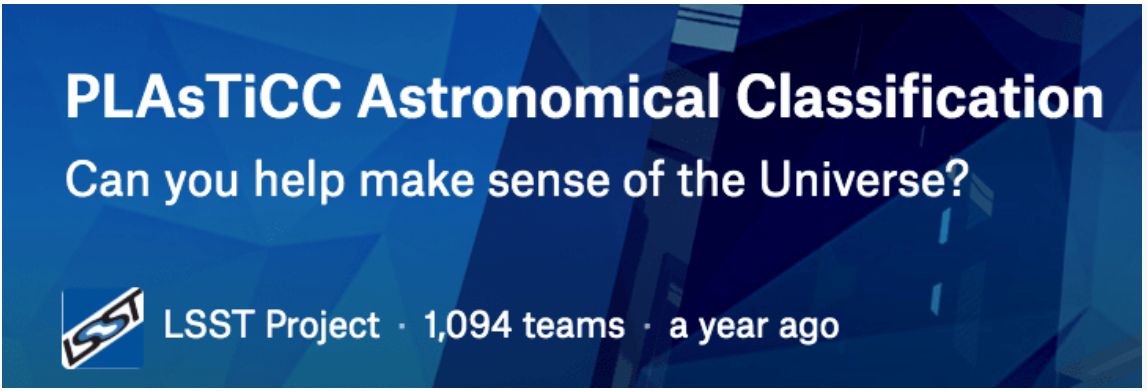
class90: Ia型超新星

https://www.kaggle.com/kenmatsu4/feature-engineering-with-gaussian-process

Gaussian Process適用後

リーダーボードの例：

このスコアは、どのくらい外れているか
を表すので、低い方が良い



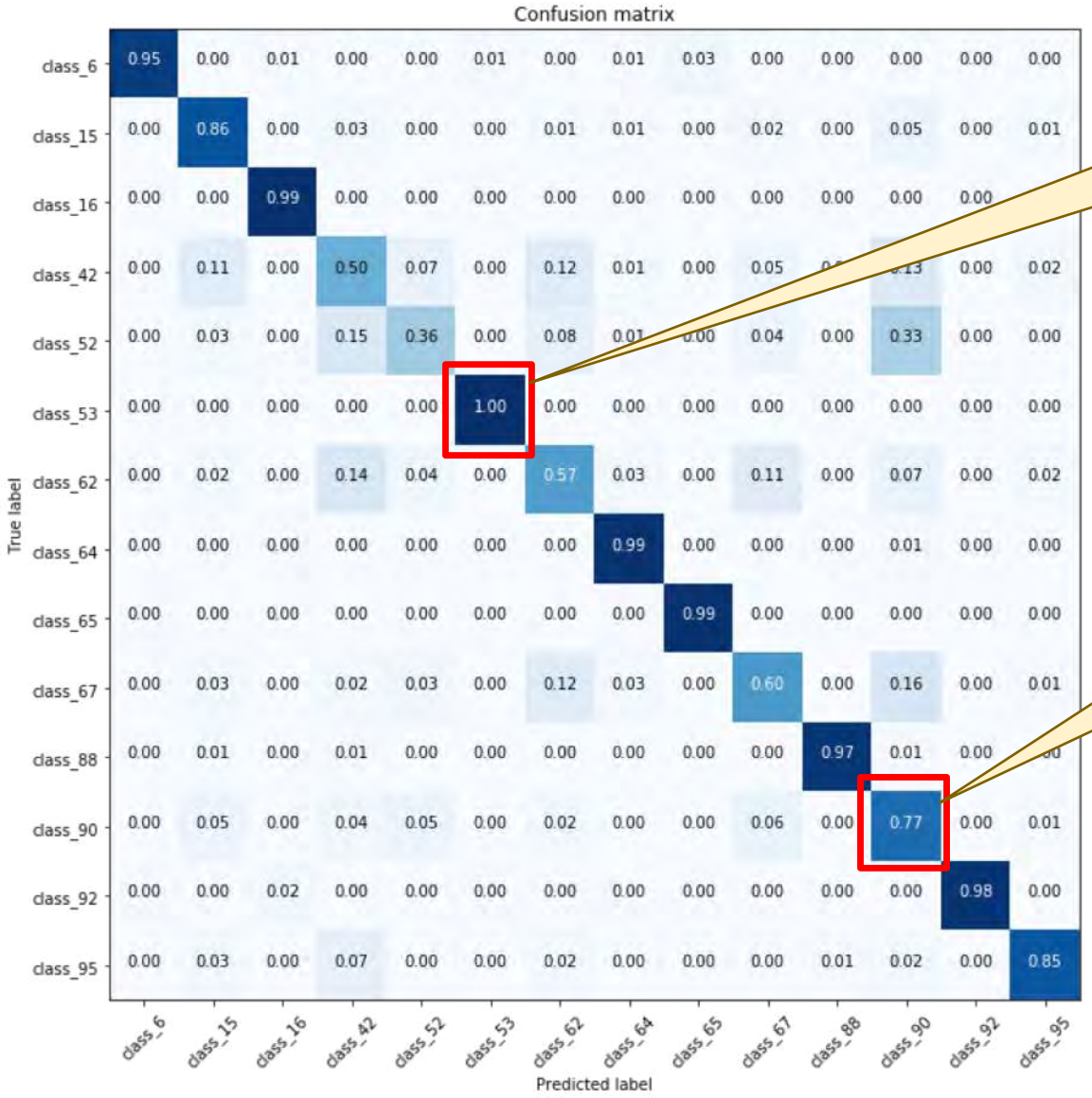
#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Kyle Boone			0.68503	104	1y
2	▲ 2	Mike & Silogram			0.69933	176	1y
3	▼ 1	Major Tom			0.70016	366	1y
4	▼ 1	AhmetErdem			0.70423	233	1y
5	—	SKZ Lost in Translation			0.75229	337	1y
6	▲ 2	Stefan Stefanov			0.80173	28	1y
7	▲ 3	hklee			0.80836	63	1y
8	▼ 1	rapids.ai			0.80905	133	1y
9	▼ 3	Three Musketeers			0.81312	313	1y
10	▲ 3	J&J			0.81901	246	1y
11	▼ 2	SimonChen			0.82247	131	1y
12	▼ 1	Go Spartans!			0.82652	148	1y
13	▼ 1	Day meets Night			0.82691	164	1y
14	▲ 6	Belinda Trotta			0.84070	105	1y
15	▼ 1	Great Square of Pegasus			0.84431	365	1y
16	▼ 1	SPACE curry			0.84620	219	1y
17	▼ 1	fakePLAsTiCCtrees			0.84653	220	1y
18	▼ 1	Is there life on Mars?			0.85180	200	1y
19	▼ 1	Stardust Crusaders★≡			0.85180	316	1y
20	▼ 1	MACHO			0.85652	131	1y
21	▲ 1	Ground Control To			0.86777	288	1y

https://www.kaggle.com/c/PLAsTiCC-2018/leaderboard

:DeNA

どれくらい精度よく当てることができたか

5位のチームの途中経過の精度表。



class53: 脈動変光星 は
ほぼ100%正解できる

class90: Ia型超新星は
77%程度で正解できる

https://www.kaggle.com/c/PLAsTiCC-2018/discussion/70669

Kaggleのコンテストのその他の例

最近開催されたコンテストから5つのコンテストを紹介。分析コンテストのほとんどは教師ありの枠組み。

	データ	概要	データ種類	参加チーム数
Home Credit Default Risk	金融機関のローン申込み	消費者ローンの申込書と、過去の返済履歴から申込者が返済不能に陥るかどうかを予測する。	テーブルデータ	7190 teams
Elo Merchant Category Recommendation	ECサイトの購買履歴	ユーザーの購買行動データから、Royalty Scoreという値を予測する。	テーブルデータ	4127 teams
iMet Collection 2019 - FGVC6	美術品の写真	NTのMetropolitan Museum of Art所蔵の美術品が美術品の種類を示すラベルを予測する。	画像	446 teams
TGS Salt Identification Challenge	地質調査画像	地質調査画像からどの領域が、塩が固まっている領域を特定する。	画像	3229 teams
Quora Insincere Questions Classification	WebのQAサイト質問/回答文	質問文が、差別的・攻撃的など不適切な文章ではないかを分類。	自然言語	4037 teams

機械学習を学ぶ/Kaggleで活躍するために必要な知識

機械学習を学ぶ、分析コンテストで活躍するには複数の分野の知識を組み合わせることが重要。

プログラミング	✓ 機械学習やKaggleでポピュラーなプログラミング言語はPython / Rです。基礎的なプログラミング力をつけることでデータハンドリング能力を高め、思いつく工夫を実現しやすくなります。
数学	✓ 基礎的な微分積分・線形代数を習得しておくことで、機械学習モデルがどのような計算を行なっているかの理解に役立ちます。データハンドリングの多くは行列の演算として考えられます。
確率・統計	✓ 機械学習は統計的学習と呼ばれることもありますが、データよりパターンを見出しそれにより目的の数値やカテゴリを予測します。予測結果は確率的に表現されることが多く、確率・統計の知識が重要です。
英語	✓ Kaggleなどの分析コンテストは日本だけでなく世界中からDSが集まっており、共通語は英語です。英語が読めることに加え、書くことができるとKaggle場でDS同士でコミュニケーションすることができます。Meet upなども開催され直接あってコミュニケーションする場もあり、活躍の場は広がります。

アジェンダ

1. 自己紹介 & 今までのキャリア
2. DeNA紹介
3. データ分析コンテストの仕組み
4. 事業で活躍するKaggler
5. おわりに

AIサービス開発の実際

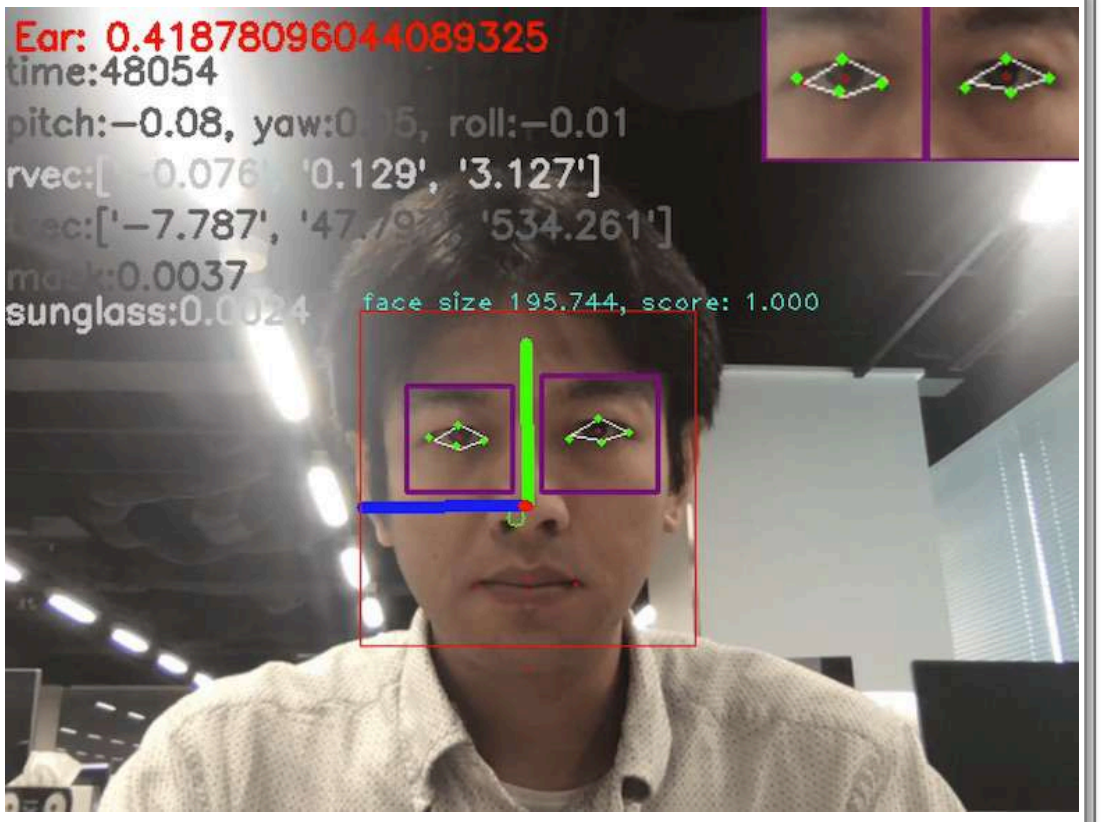
～ DRIVE CHARTを例にとって ～

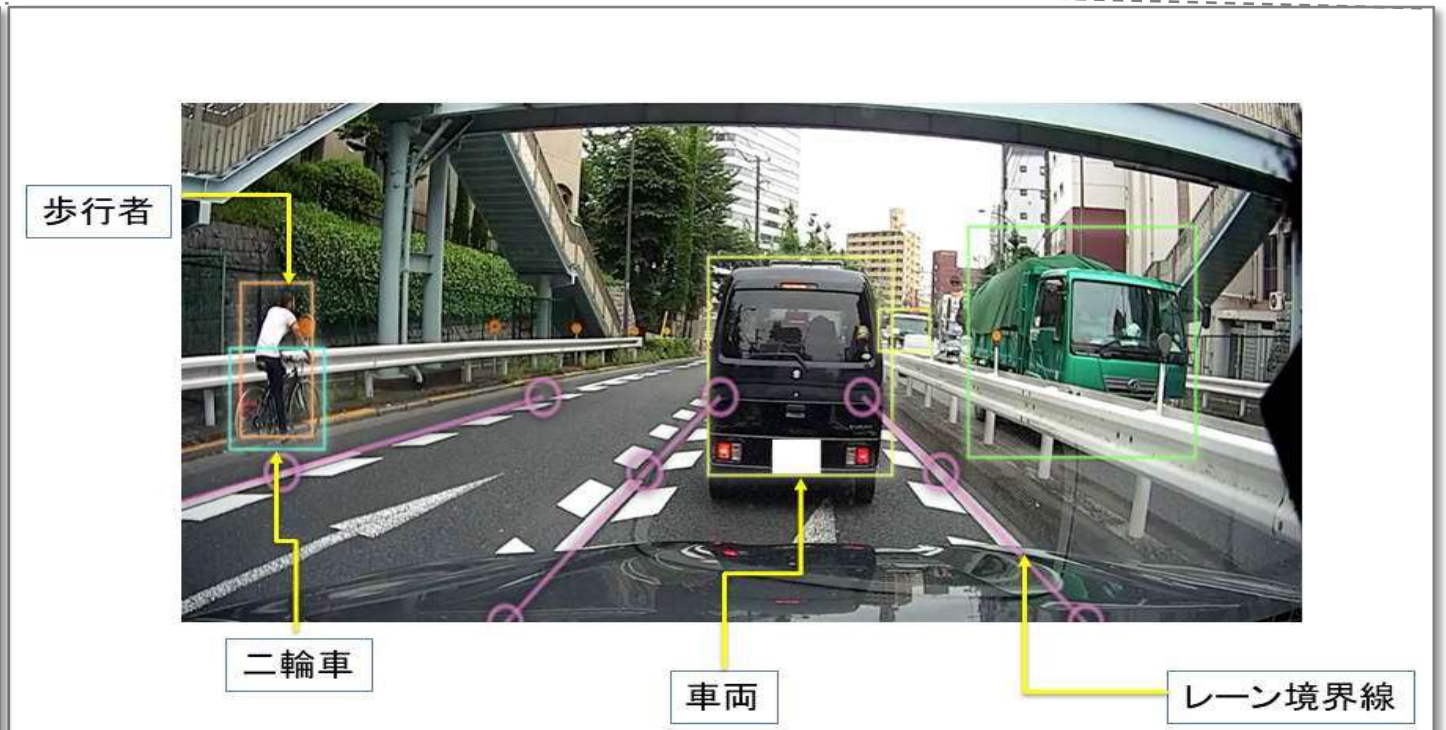
DRIVE CHARTとは https://drive-chart.com/

交通事故の削減を目指し、安全を脅かす様なドライバーの運転の癖や行動をAIが検出し、運転行動の改善へと導くサービス。

DRIVE CHARTのにおけるDeep Learning画像認識の活用

内側向きカメラを用いた脇見検出、外側向きカメラを用いた車間距離不足検出に活用しています。



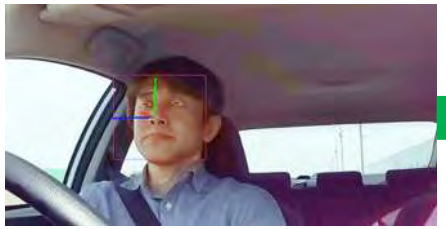


システム構成

エッジデバイスで得られたデータやAIの推測結果をサーバーに集め、イベント検出を行い顧客にレポート提供。

エッジデバイス

外カメ画像



内カメ画像

深層学習モデル

Object Detection 結果

Lane Detection 結果

Face Landmark 検出結果

加速度センサ

GPS

クラウドサーバー

データベース

イベント検出モデル

- 急加速
- 急減速
- 急ハンドル
- 一時不停止
- 速度超過
- 車間距離不足
- 脇見

クライアント

レポート表示

DRIVE CHART チーム体制

DRIVE CHARTのサービスは様々な役割を持った人たちが力を合わせて構築・運営している。

AI チーム

AI 研究開発エンジニア

コンピュータビジョンの技術領域で高い専門性を活かし、エッジデバイスに搭載するDeep Leaningモデルを構築。

データサイエンティスト

加速度・速度、位置情報などのセンサーデータや、Deep Learningの予測結果を統合し、危険シーン検知モデルを構築。

機械学習エンジニア

機械学習モデルの開発検証サイクルを支える仕組みづくりや、データパイプライン整備、データベース整備など大規模データのハンドリングが可能な仕組みを構築。

事業 チーム

プロダクトマネージャ

サービスのコンセプトやRoad Mapなどの方向性を定めたり、プロダクトの仕様を決める。

エッジデバイスエンジニア

車載デバイスに搭載するソフトウェアを限られたリソースを最大限活用して開発/検証を行う。CVモデルの組み込みやデータ生成・アップロードの仕組みなど

サーバサイドエンジニア

車載デバイスから上がってくるデータをDBに保存、AI処理に回し結果をレポート表示する仕組みを構築。

事業開発・推進

安定してサービス稼働できるよう顧客サポートを行なったり、顧客の要望からサービスの改善を実施したり、ビジネススキームの検討などを行う。

データサイエンティストが持つスキル

データサイエンティストが抑えるべき領域は多岐にわたる。下記は松井の考える必要スキル。

数理知識
(機械学習・統計学)

- ✓ 機械学習（予測モデル構築）
- ✓ 統計学（データの生成メカニズムを解き明かす）
- ✓ 最適化

エンジニアリング

- ✓ プログラミング（Pythonなど）
- ✓ データベース
- ✓ クラウドサービスの知識
- ✓ データ前処理、データパイプライン構築

データインサイト
EDA ※

- ✓ データ可視化、探索
- ✓ 特徴量エンジニアリング
- ✓ データの特性に関する知識

レポーティング

- ✓ わかりやすい説明力
- ✓ ロジカルシンキング
- ✓ 資料作成力（パワーポイントなど）

仮説出し
示唆出し

- ✓ 想像力
- ✓ 置かれた状況を把握する力
- ✓ データからパターンを読み解く力

業務推進力

- ✓ プロジェクト関係者とのコミュニケーション力
- ✓ 段取り力（スケジューリング）

ドメイン知識
（分析対象に対する知識）

- ✓ 関わるプロジェクトに関するドメイン知識

※ Exploratory Data Analysis：探索的データ解析

アジェンダ

1. 自己紹介 & 今までのキャリア
2. DeNA紹介
3. データ分析コンテストの仕組み
4. 事業で活躍するKaggler
5. おわりに

おわりに

- データサイエンスは、数理的な知識、プログラミングを活用して仕事をすることができるエキサイティングな分野。
- データ分析コンテストの利点
  - ✓ 世界には高校生でGrandmasterになった人も。年齢も職業もバックグラウンド関係なく参加可能で、力試しができる。
  - ✓ 世界トップレベルの知見を具体的に知ることができる。
  - ✓ Globalにコミュニケーションを取ることができる。
- データ分析、機械学習を活用したサービス・製品が世の中に提供されています。

EOF