

Tomo-e Gozenとビッグデータと機械学習

－ 観測システムへの実装の観点から －

酒向重行（東京大学大学院理学系研究科附属天文学教育研究センター）,

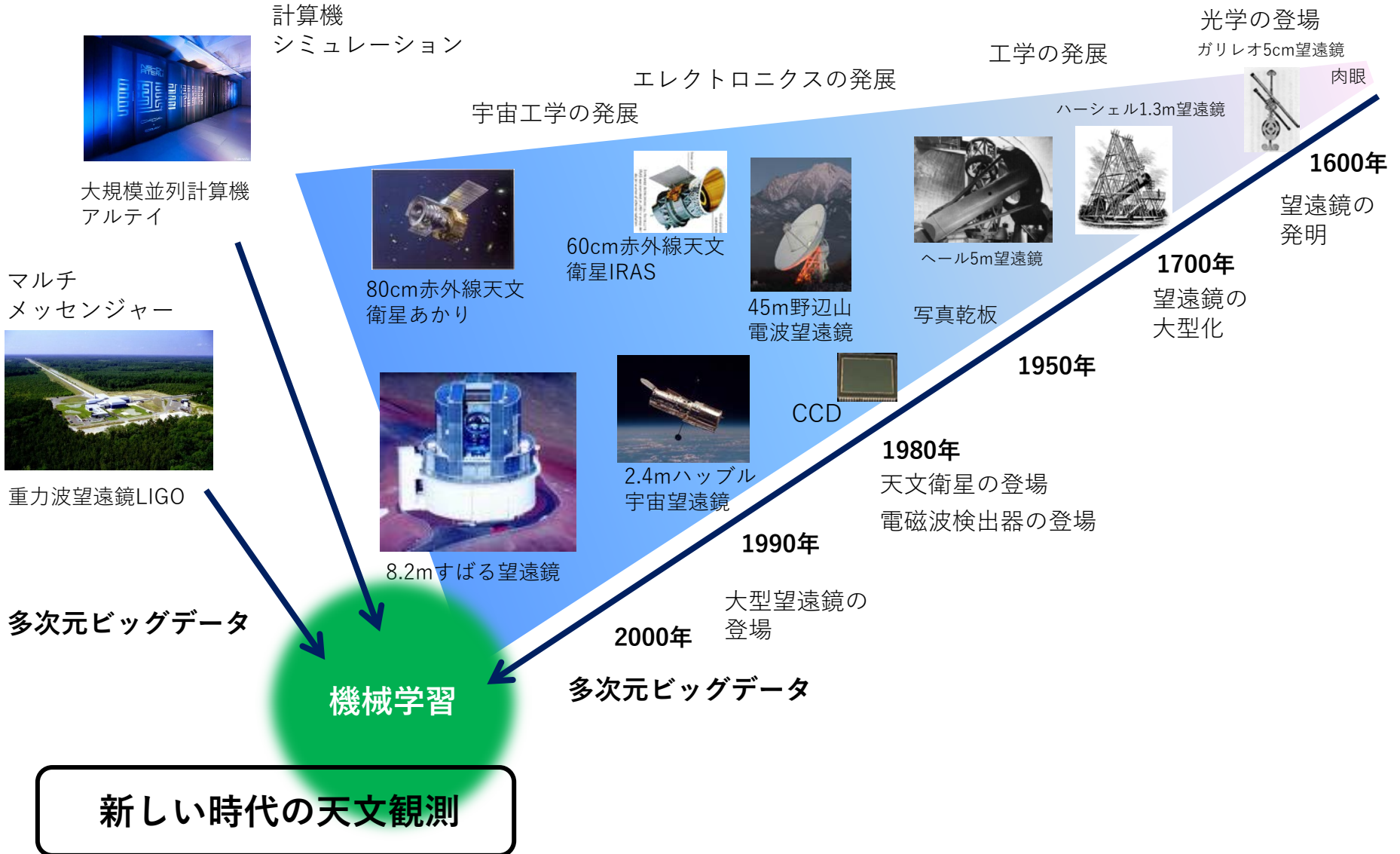
Tomo-e Gozen コラボレーション



2020/9/15 15:35-

2020年度光赤天連シンポジウム データ解析の新展開2020

天文観測技術の歩み



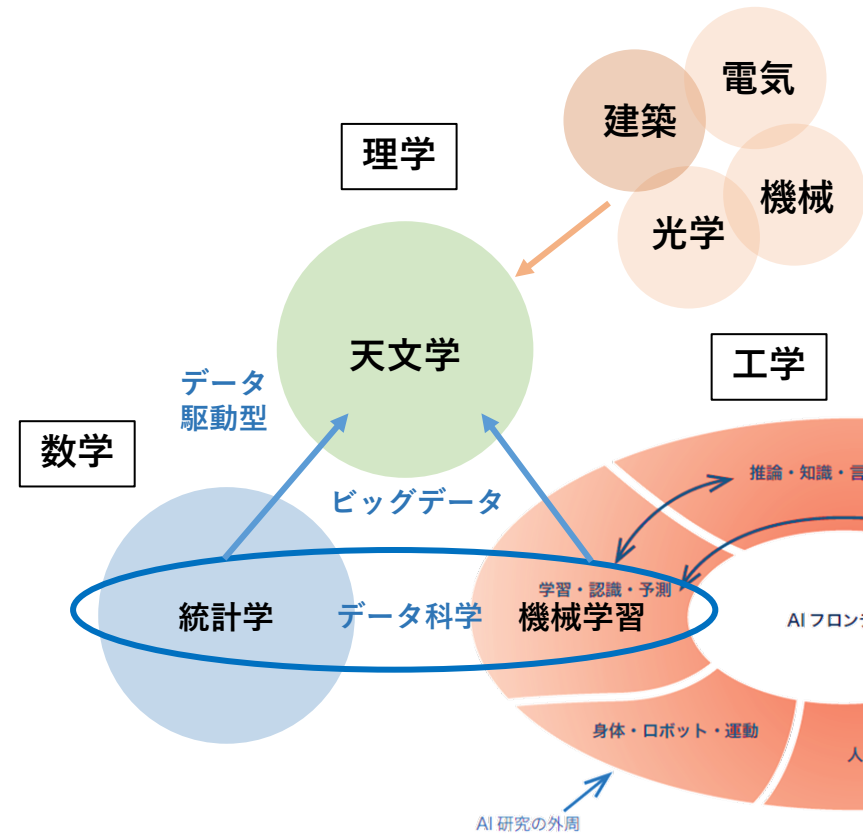
天文学における機械学習

用途

- 予測モデルの構築（特に経験モデル）
- 多次元空間からの隠れた情報の抽出
- 情報の分類、認識
- 異常検知、情報補間
- ノイズ除去、偽検出の除去

機械学習を前提とした観測システムの構築が大切

- 最初から観測システムへの導入を前提とする
- 必要なデータを取得、不必要なデータは取得しない
- ビッグデータの獲得
- データ駆動型の導入



Tomo-e Gozen

データ駆動型科学の理念に立脚した宇宙観測システム

宇宙シミュレーション + 広視野動画センサ

空のビッグデータ

ビッグデータ, 機械学習, 最適化,
自動観測, 高速ネットワーク

即時共有

分類、偽情報の除去
アラート生成
的確な判断

即時解析



世界初の天文用広視野動画カメラ

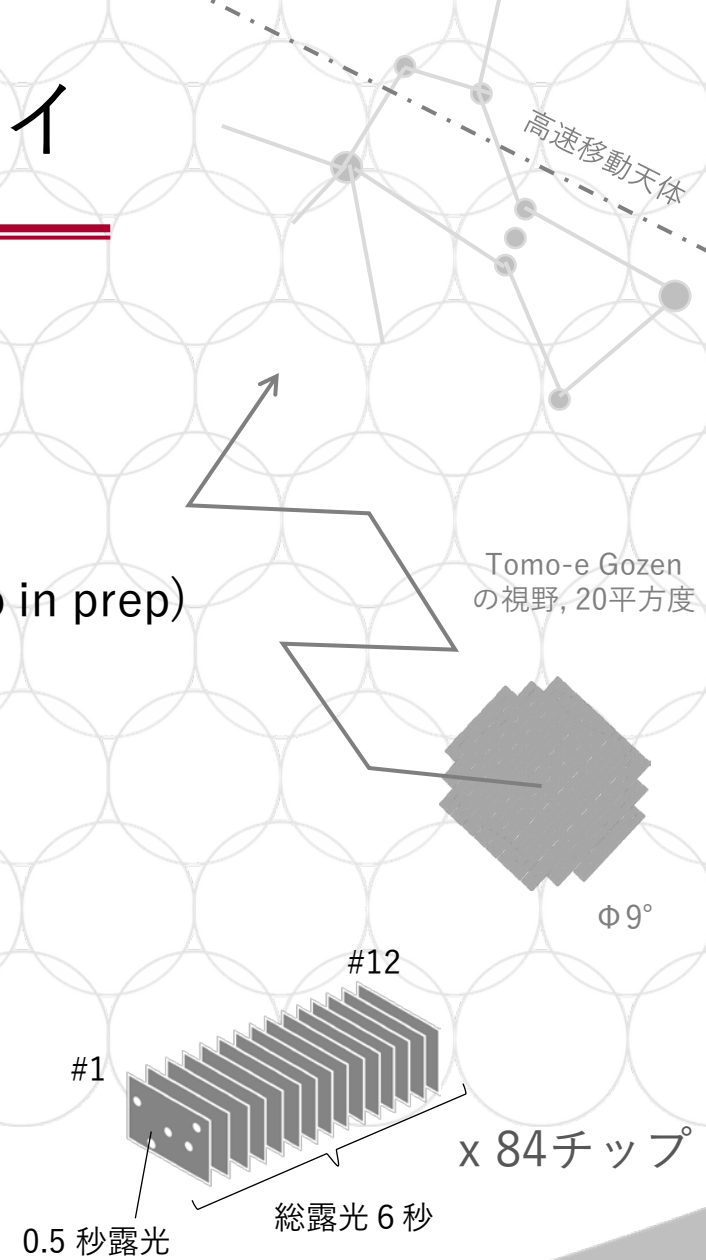
Tomo-e Gozen

- 木曾105cmシュミット望遠鏡
- 視野 20平方度
- 84台の2k x 1k CMOSセンサ
- 2フレーム/秒の動画
- 毎夜30TBのビッグデータ
- 可視光単色 (g+r相当)
- 運用期間 >10年



Tomo-e Gozen 広域動画サーベイ

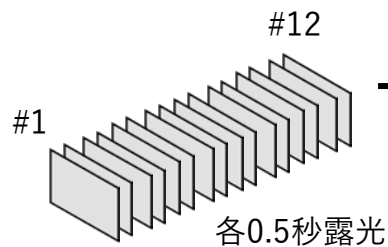
- 20平方度の動画, 0.5秒 x 12フレーム
- <17等級@各フレーム, <18等星@スタック画像
- 晴天夜のほぼ毎日 (~100 夜/年)
- サーベイ経路の自動スケジューリング (J. P. Pedroso in prep)
- 2019年10 - 12月
 - 7,000平方度 (高度 > 35°), 2時間頻度, 3-5回/夜
- 2020年1 - 6月
 - 7,000平方度 (高度 > 35°), 1回
 - 2,000平方度 (深夜南中領域), 0.5時間頻度, ~8回/夜



Hoursタイムスケールの突発天体の検出

Tomo-eシステム ← → サイエンスチーム

Tomo-e Gozen動画データ



計6秒露光
画像

キャリブレーション

Pan-STARRS画像 (PS1,r-band)

差分画像

点源の検出
 10^8 候補/夜

引き算前後の
カットアウト画像

数分 - 数10分

突発現象カタログ

候補天体の
測光データ

機械学習モデル
で分類

既存カタログ
で分類

10^4 候補/夜

偽信号を除去 (人工的なノイズ,
シーイングダンス, 宇宙線など)

ウェブ
インター
フェース

追観測用
アラート

超新星

AGN

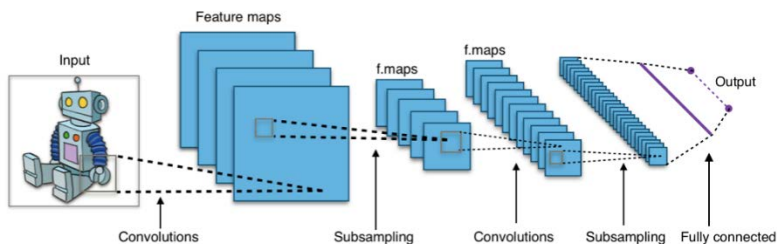
星フレア

太陽系
小天体

AGN, 突発天体, 太陽系小天体の多波長カタログ (optical, UV, IR: SDSS, PS1, GALEX, WISE, AKARI...; X-ray: ROSAT, XMM-Newton, Chandra...; radio: FIRST, NVSS, TGSS; SDSS(BAL) quasars, TNS transient, MPCChecker...)

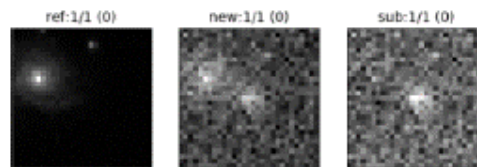
畳み込みニューラルネットワーク(CNN)による超新星の検出

浜崎ら (日本天文学会2020年春季年会, Z413a)



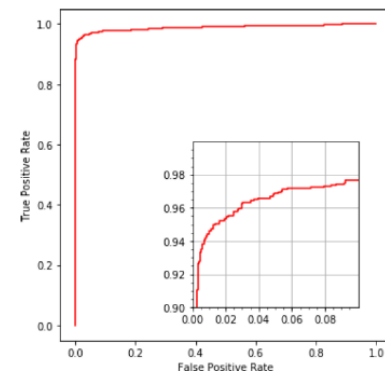
畳み込みニューラルネットワークの模式図 (Wikipediaより)

教師データ(画像3種+超新星かどうか)



参照画像 (過去) Tomoe 画像 差分画像

- 29 x 29画素, 3チャンネル
- シミュレーション(疑似)画像を作成
- 学習データ 本物14,000, 偽物14,000



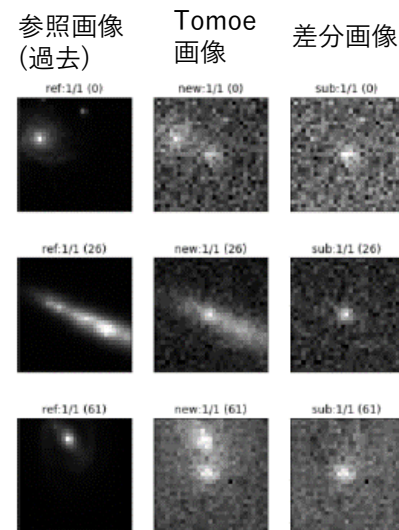
学習したモデルのROC曲線

CNNモデル

- 畳み込み (filter 5x5, 32種)
- プーリング (2x2)
- 畳み込み (filter 3x3, 64種)
- プーリング (2x2)
- ドロップアウト
- 全結合NN (1600入力)
- 128 (活性化関数ReLU)
- 64 (活性化関数ReLU)
- 1 (活性化関数シグモイド)

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 25, 25, 32)	2432
max_pooling2d_1 (MaxPooling2)	(None, 12, 12, 32)	0
conv2d_2 (Conv2D)	(None, 10, 10, 64)	18496
max_pooling2d_2 (MaxPooling2)	(None, 5, 5, 64)	0
dropout_1 (Dropout)	(None, 5, 5, 64)	0
flatten_1 (Flatten)	(None, 1600)	0
dense_1 (Dense)	(None, 128)	204928
dense_2 (Dense)	(None, 64)	8256
dense_3 (Dense)	(None, 1)	65

出力: 超新星である確率(0-1)



CNNモデルでTomoe-eデータに検出した超新星の例

超新星候補の検出の実績

†典型的に半数程度が超新星の候補

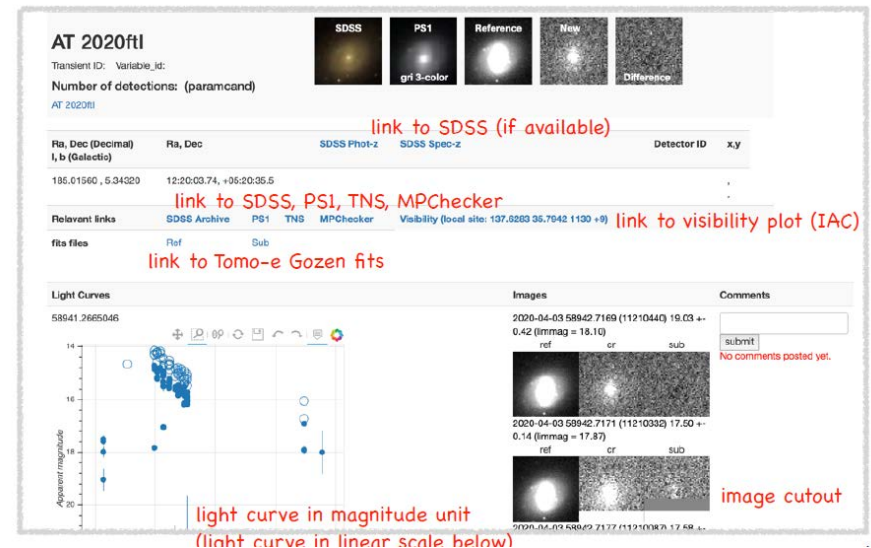
Transient Name Server (TNS)カタログに登録されている突発天体†と比較

- 2019/4/6 – 12/31にTNSに登録された突発天体 (Dec > -3 deg) : 1,600天体



全体: TNS検出 1,600天体

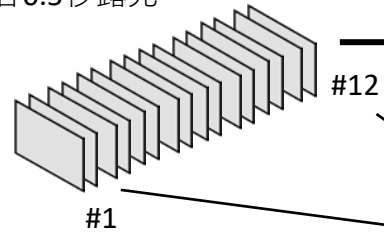
Tomo-e Gozen突発天体確認用ウェブビューア



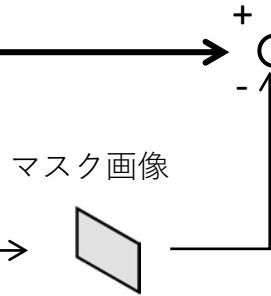
高速移動天体(>2"/秒)の検出

Tomo-e Gozen動画データ

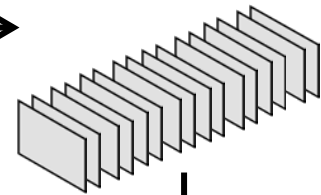
各0.5秒露光



静定天体をマスク

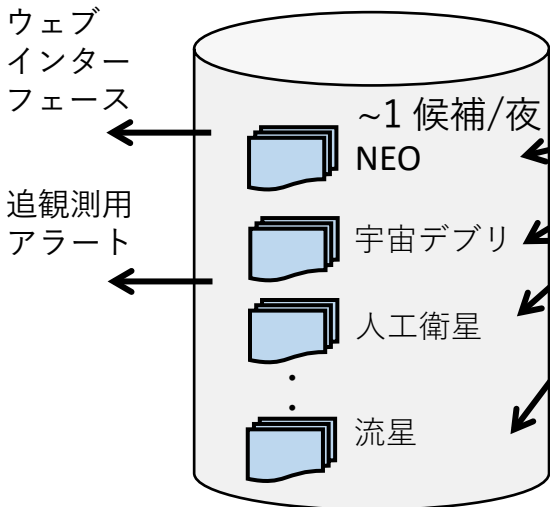


移動天体抽出動画



数分 - 数10分

高速移動天体カタログ



既存カタログで分類



10³ 候補/夜

機械学習モデルで分類



偽信号を除去 (人工的なノイズ, シーイングダンス, 宇宙線など)

人工衛星, 宇宙デブリ, NEOカタログ

点源の検出
点源間の同定
10⁶ 候補/夜

静止画・動画の
特徴量

サイエンスチーム ←

→ Tomo-eシステム

ランダムフォレストによる高速移動天体の学習

決定木

- ノード間の情報利得(不純度の差)を最大化する順
- パラメータ化した特徴量を使って学習

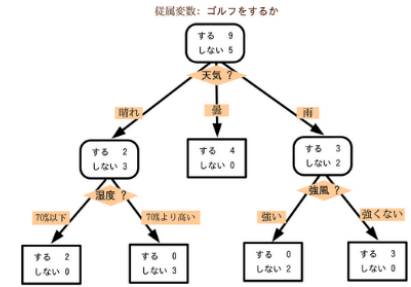
- ジニ不純度 $I_G(t)$

$$I_G(t) = 1 - \sum_c p(i|t)^2 \quad \text{ただし } p(i|t) = \frac{n_i}{n}$$

t: ノード
c: class (小惑星、人工衛星)

- 情報エントロピー $I_H(t)$

$$I_H(t) = - \sum_c p(i|t) \log_2 p(i|t)$$



不純度の例

決定木の例 (Wikipedia)

ランダムフォレスト

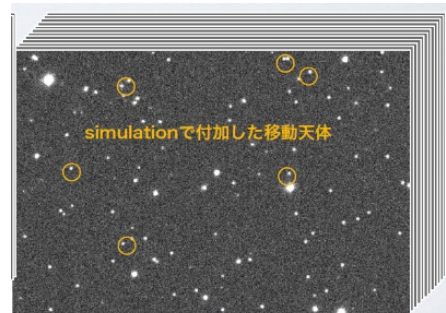
- 決定木を複数使用することで過学習を抑制
- データと特徴量の種類をランダムサンプリングし決定木を多数生成
- 多数の決定木の結果を集約して総合的な判断 (アンサンブル学習)

学習

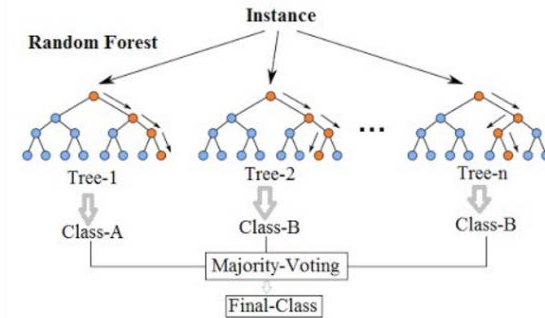
- Tomo-e Gozenで取得した動画データに疑似移動天体(シミュレーション)を埋め込み、移動天体検出ソフトにて検出と特徴量の抽出
- 学習に使った特徴量

- 背景カウント
 - 天体形状
 - 移動速度
 - 動画内検出数
- } 静止画的特徴量
- } 動画的特徴量

など計19コ



実際の動画データに疑似移動天体を付加(シミュレーション) これらを教師データのTrueとする



ランダムフォレストの概要 (Wikipedia)

高速移動天体確認用ウェブビューア

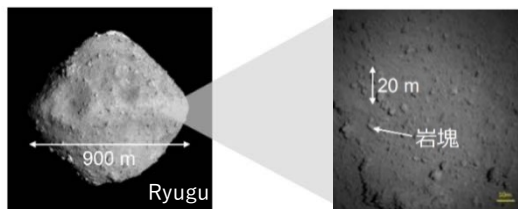
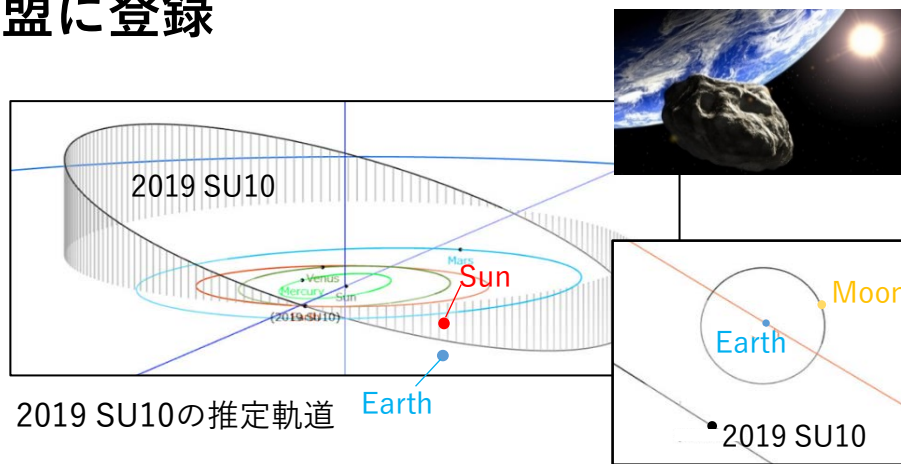
NO.	DATE(UTC), OBJECT, FITS	MATCHING	IMAGE	LIGHT CURVE	SCORE	RA, DEC	V_MEAN ["/sec]
13	2020-01-02T15:07:37.847 20amkwy_08011+20691 TMQ3202001020021826813.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.983	08h00m41.1176s +20d41m27.3921s	5.35
75	2020-01-02T12:08:03.495 20amkpz_02342+05475 TMQ3202001020021808713.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.982	02h20m29.7541s +05d28m20.9898s	15.18
101	2020-01-02T11:25:29.630 20amklu_04576-05758 TMQ1202001020021797841.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.981	04h34m33.9624s -05d45m30.1496s	7.65
95	2020-01-02T11:27:25.319 20amkma_04592-05750 TMQ1202001020021798431.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.979	04h35m32.1233s -05d45m02.3742s	7.54
4	2020-01-02T15:45:25.683 20amlaj_07983+19973 TMQ4202001020021835714.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.968	07h58m58.351s +19d58m26.8782s	6.44
110	2020-01-02T10:17:42.845 20amkei_04612-07426 TMQ3202001020021778411.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.964	04h36m44.5036s -07d25m26.4031s	10.36
116	2020-01-02T10:09:42.704 20amkdl_03428+04611 TMQ3202001020021776114.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.958	03h25m38.5353s +04d36m29.5043s	14.63
120	2020-01-02T10:05:14.537 20amkcy_04489-03946 TMQ4202001020021774816.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.956	04h29m19.8435s -03d56m45.5029s	12.36
1	2020-01-02T16:02:38.498 20amlcg_05309+12310 TMQ3202001020021840614.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.944	05h18m33.8163s +12d18m34.7262s	6.8

20amknl_05285-04523 TMQ2202001020021802932.fits Tracklet Search Circle fit	Satellite NEO NEOCP			0.895	05h17m06.9861s -04d31m28.2281s	7.28
2020-01-02T15:00:34.544 20amkwf_05498-04235 TMQ1202001020021824944.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.892	05h29m52.3445s -04d14m16.6638s	5.73
2020-01-02T15:13:40.957 20amloh_07252+31691 TMQ2202001020021827713.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.887	07h15m06.8848s +31d41m33.1599s	6.25
2020-01-02T12:02:47.375 20amkpk_02634+01881 TMQ1202001020021807232.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.887	02h38m00.6009s +01d52m41.7195s	14.86
2020-01-02T12:10:29.257 20amkqg_02473-05748 TMQ2202001020021809432.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.881	02h28m21.6532s -05d44m55.3739s	15.15
2020-01-02T12:09:37.519 20amkqe_02499+08385 TMQ2202001020021809242.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.881	02h29m54.2345s +08d23m05.277s	15.39
2020-01-02T12:02:11.739 20amkpi_02756+01494 TMQ1202001020021807022.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.881	02h45m23.1962s +01d29m46.7819s	10.92
2020-01-02T11:25:12.068 20amklt_04708-04014 TMQ1202001020021797723.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.881	04h42m28.2843s -04d00m57.0347s	7.13
2020-01-02T10:17:05.653 20amkeg_04622-05985 TMQ2202001020021778212.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.88	04h37m18.4347s -05d59m08.097s	7.18
2020-01-02T10:11:46.413 20amkdr_03181+02923 TMQ1202001020021776734.fits Tracklet Search Circle fit	Both Satellite NEO NEOCP			0.879	03h10m51.2492s +01d18m34.7262s	15.84

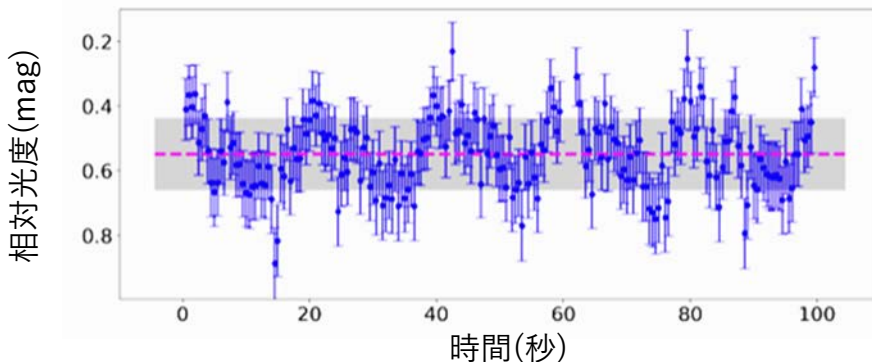
小型NEO(地球接近小惑星)の発見実績

最近1年に13個のNEOが国際天文学連盟に登録

#	名前	発見日	推定直径	地球最近距離
13	2020 QW	2020/8/17	31m	177万km
12	2020 PW2	2020/8/14	6m	28万km
11	2020 HT7	2020/4/27	15m	71万km
10	2020 HU3	2020/4/21	22m	125万km
9	2020 GY1	2020/4/2	17m	8万km
8	2020 FA2	2020/3/18	11m	54万km
7	2020 EO	2020/3/12	24m	93万km
6	2019 XL3	2019/12/15	14m	53万km
5	2019 XT2	2019/12/08	19m	54万km
4	2019 XM2	2019/12/05	17m	88万km
3	2019 VD3	2019/11/05	24m	101万km
2	2019 SU10	2019/09/25	13m	100万km
1	2019 FA	2019/03/16	6m	23万km



はやぶさ2が撮影したRyuguの表面。



10m級NEOの光度曲線 (2fps動画)

† 月距離(38万km), 静止衛星軌道(3.6万km)

サーベイ経路の自動スケジューリング

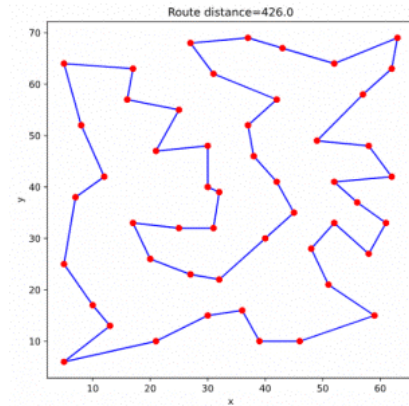
Pedroso, Ikeda, Morokuma, & Sako in prep.

目的

- Tomo-e Gozenは読み出し時間がゼロなので、効率的な経路決定が重要
- 効率的なサーベイ経路を自動決定したい
- 天候変化、ターゲット変更に対してロバストにしたい

最適化手法

- ヒューリスティック最適化モデル(発見的な方法)を採用 → 高速+精度そこそこ
- ローリングホライゾン法（数手先までを最適化を繰り返す）と、巡回セールスマン問題の解法を組み合わせる



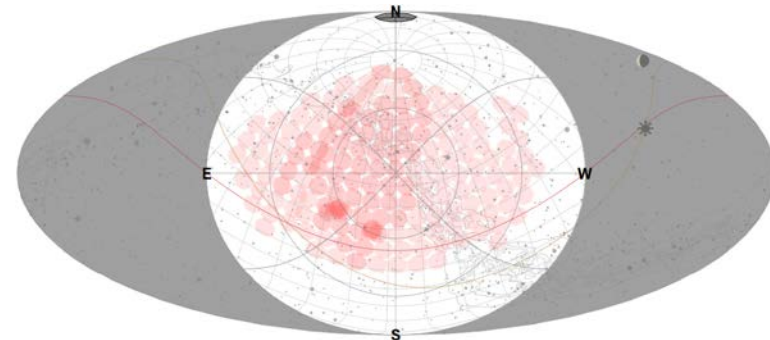
巡回セールスマン問題の解の例

最適化条件

- 特定の時間間隔($> \Delta t$)で、1晩に複数回($> N$)の訪問を実現する領域を最大化 ($N \sim 5$)

今後の課題

- 天候変化、ターゲット変更などの条件の変更の度に、再計算するよう、システムに実装
- 特定の部分領域の観測（重力波対応）



Tomo-e Gozenの広域サーベイのフットプリント

Tomo-e Gozenとビッグデータと機械学習

Tomo-e Gozenの動画ビッグデータ

- 20平方度の動画, 0.5秒 x 12フレーム, 晴天夜のほぼ毎日 (~100 夜/年)
- <17等級@各フレーム, <18等星@スタック画像

機械学習の事例

- 機械学習モデルによる高信頼 + 高速な偽信号除去
- 畳み込みニューラルネットワーク: 超新星の検出。TNSよりも早期に検出。
- ランダムフォレスト: NEO,人工天体の検出, 約1年で13個の小型NEOの発見
- 最適化アルゴリズムによるサーベイ経路の自動スケジューリング

機械学習を前提とした観測システムの構築が大切

- 最初から観測システムへの導入を前提とする
- 必要なデータを取得、不必要なデータは取得しない

天文学会に「V4 機械学習/データ科学」の登場を期待