NIIGATA UNIVERSITY

新潟大学の紹介 / 太陽データの機械学習実習

新潟大学

大学院自然科学研究科 情報工学コース

/ 工学部 知能情報システムプログラム

飯田 佑輔

March 30th, 2023

太陽研究最前線ツアー

新潟大学 大学院情報工学コース / 工学部 知能情報システムプログラム

NIIGATA UNIVERSITY

情報工学コース分野

- **コンピュータサイエンス**
- 情報ネットワーク
- マルチメディア
- 情報セキュリティ

March 30th, 2023

太陽研究最前線ツアー

自己紹介

NIIGATA UNIVERSITY

名前: Yusuke Iida (飯田 佑輔)

経歴:

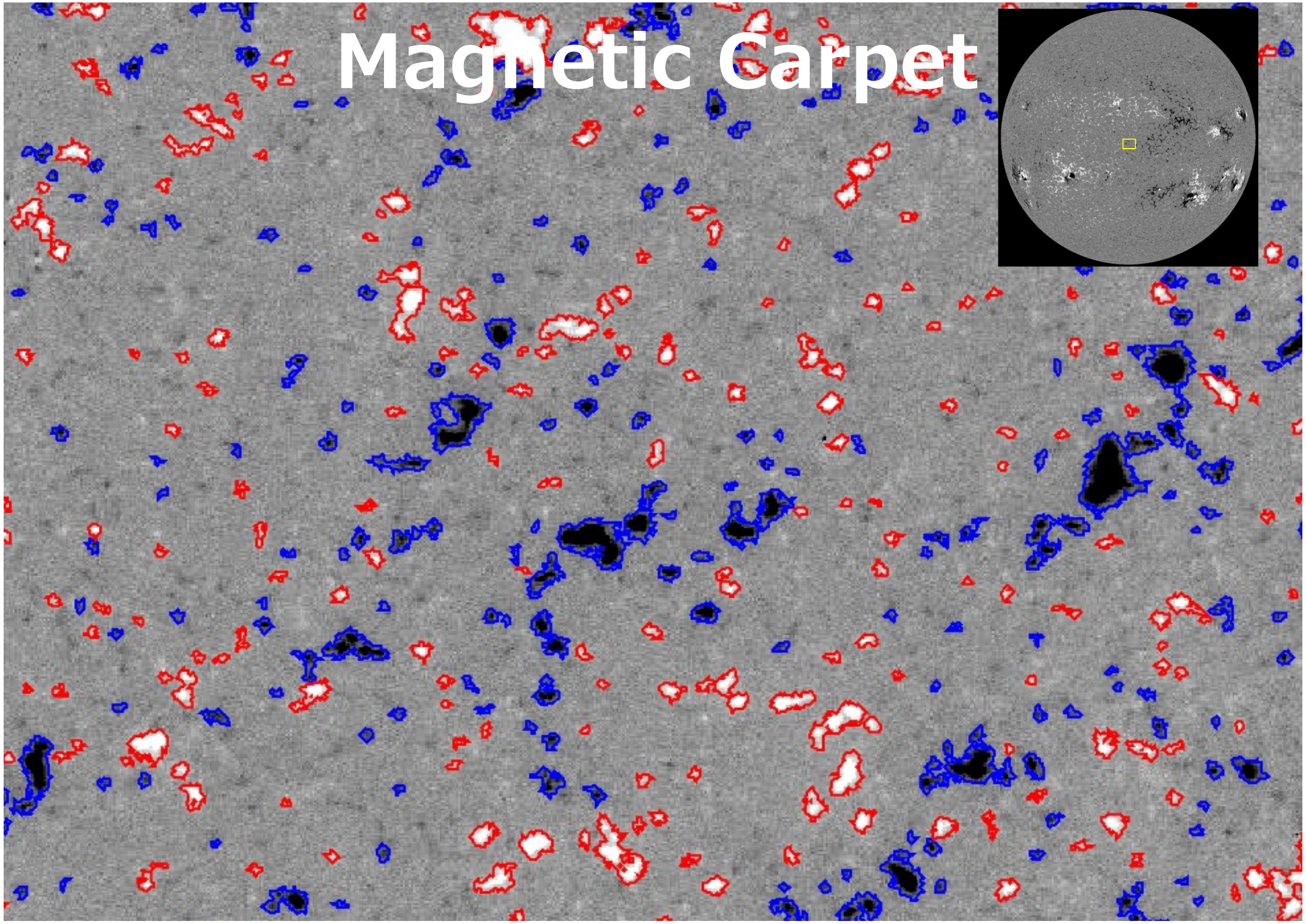
1985	三重県鈴鹿市生まれ
2003 - 2012	東京大学 理学部/地球惑星科学専攻 「**太陽静穏領域磁場の形成・維持機構**」で学位
2012 - 2013	JSPS特別研究員(PD)@東大
2013 - 2016	プロジェクト研究員@JAXA/ISAS
2016 - 2019	助手@関西学院大学/理工学部物理学科
2019 -	准教授@**新潟大学/工学部知能情報システムプログラム**

所属学会:

日本天文学会，JpGU，電子情報通信学会，人工知能学会，IEEE etc.

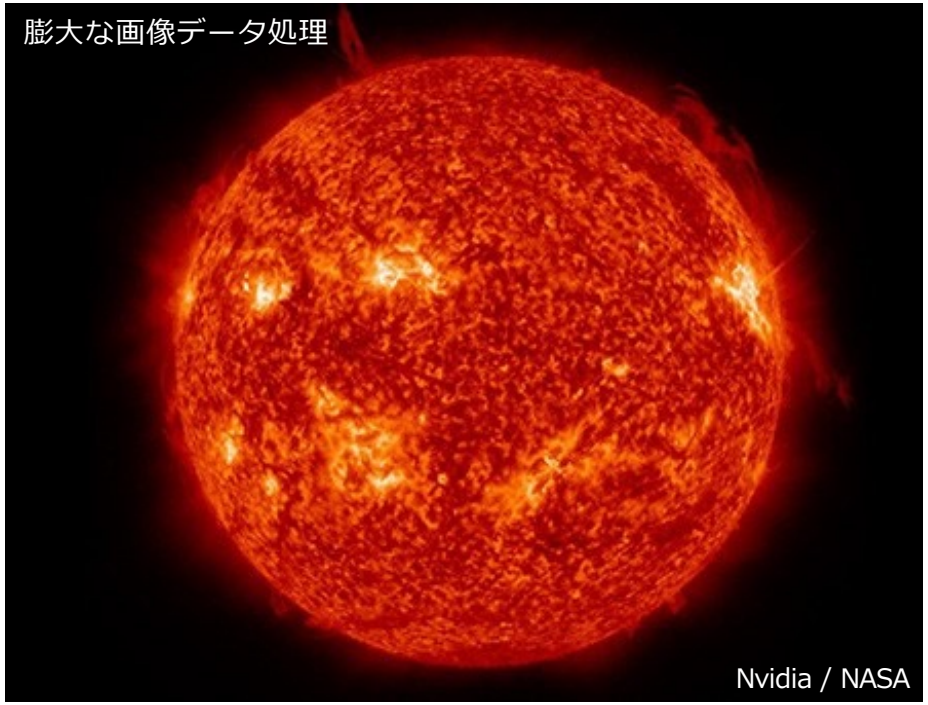
研究室の研究テーマ: 深層学習モデル構築が多い

宇宙天気/太陽	～４０% **フレア予測**，コロナホール予測，**黒点成長予測**
画像/機械学習の融合研究	～４０% 宇宙分野（銀河，光学系），農学（圃場，杉，病理診断），地球科学（氷河湖，インド洋），医学（人肌，再生医療）etc.
機械学習応用	～２０% ラベルエラー，**関数同定**


Magnetic Carpet

天文 x 情報

NIIGATA UNIVERSITY



IEEE Bigdata 2019

Challenges with Extreme Class-Imbalance and Temporal Coherence: A Study on Solar Flare Data

Azim Ahmadzadeh
dept. Computer Science
Georgia State University
Atlanta, GA, USA
aahmadzadeh1@cs.gsu.edu

Maxwell Hostetter
dept. Computer Science
Georgia State University
Atlanta, GA, USA
mhostetter1@cs.gsu.edu

Berkay Aydin
dept. Computer Science
Georgia State University
Atlanta, GA, USA
baydin2@cs.gsu.edu

Manolis K. Georgoulis
RCAAM of the Academy of Athens,
Athens, Greece
manolis.georgoulis@phy-astr.gsu.edu

Dustin J. Kempton
dept. Computer Science
Georgia State University
Atlanta, GA, USA
dkempton1@cs.gsu.edu

Sushant S. Mahajan
dept. Physics & Astronomy
Georgia State University
Atlanta, GA, USA
mahajan@astro.gsu.edu

Rafal A. Angryk
dept. Computer Science
Georgia State University
Atlanta, GA, USA
angryk@cs.gsu.edu

***Abstract*—In analyses of rare-events, regardless of the domain of application, class-imbalance issue is intrinsic. Although the challenges are known to data experts, their explicit impact on the analytic and the decisions made based on the findings are often overlooked. This is in particular prevalent in interdisciplinary research where the theoretical aspects are sometimes overshadowed by the challenges of the application. To show-case these undesirable impacts, we conduct a series of experiments on a recently created benchmark data, named Space Weather ANalytics for Solar Flares (SWAN-SF). This is a multivariate time series dataset of magnetic parameters of active regions. As a remedy for the imbalance issue, we study the impact of data manipulation (undersampling and oversampling) and model manipulation (using class weights). Furthermore, we bring to focus the auto-correlation of time series that is inherited from the use of sliding window for monitoring flares' history. Temporal coherence, as we call this phenomenon, invalidates the randomness assumption, thus impacting all sampling practices including different cross-validation techniques. We illustrate how failing to notice this concept could give an artificial boost in the forecast performance and result in misleading findings. Throughout this study we utilized Support Vector Machine as a classifier, and True Skill Statistics as a verification metric for comparison of experiments. We conclude our work by specifying the correct practice in each case, and we hope that this study could benefit researchers in other domains where time series of rare events of are interest.**

***Index Terms*—class imbalance, sampling, time series, flare forecast**

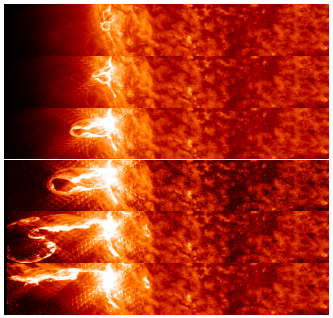


Fig. 1. Snapshots of an X-class flare, peaking at 7:49 p.m. EST on Feb. 24, 2014, observed by NASA's Solar Dynamics Observatory, in the 304Å wavelength channel. (Images source: https://helioviewer.org/)

I. Introduction

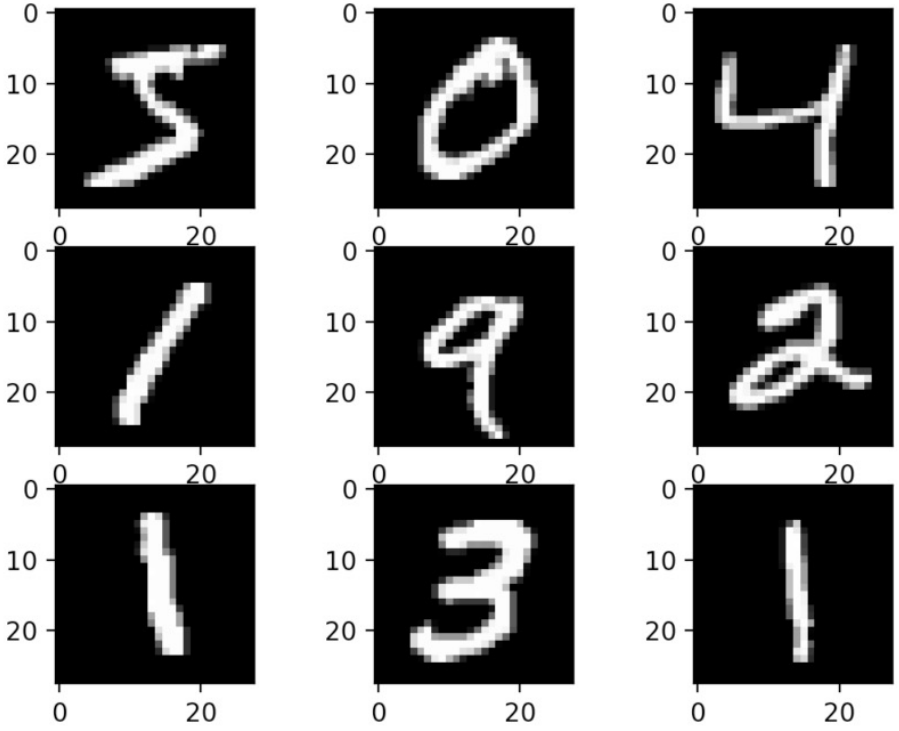
To gain valuable insights or robust predictive performance from data, we must first ensure the integrity of our data. Beyond data collection, this involves data-cleaning. It requires a thorough investigation by the experts of the domain and data scientists to produce a reliable dataset. Nonetheless, there are some challenges which are inherited from the subject under study due to unique characteristics of the data which should be identified, understood and dealt with appropriately. Class-imbalance issue is one of the main problems of this kind, which is present in many natural or other nonlinear dynamical systems. This is often due to the nature of the events, not the data collection process.

Class-imbalance is a common problem, with many potential remedies. Some of these remedies are common and well-known, but can still be misapplied. This is particularly true when the primary objective is not machine learning per se but the testing and scrutiny of domain-specific theories. The complexity of the problem at hand and the absence of data experts very often underestimate the needed level of care,

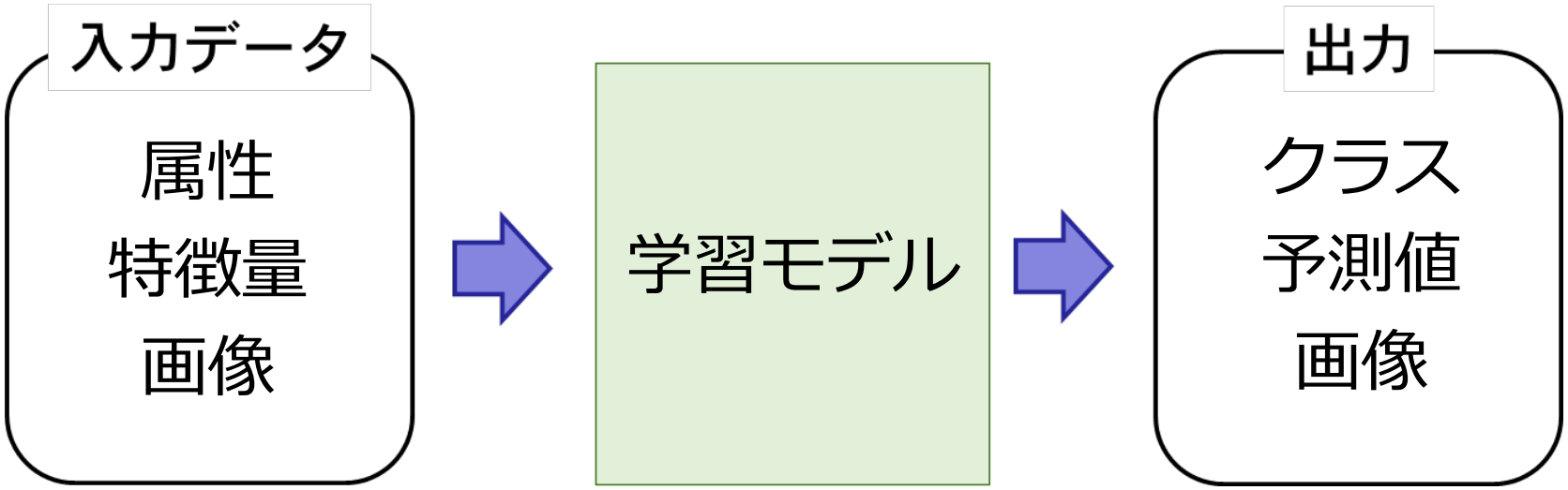
機械学習 / 深層学習

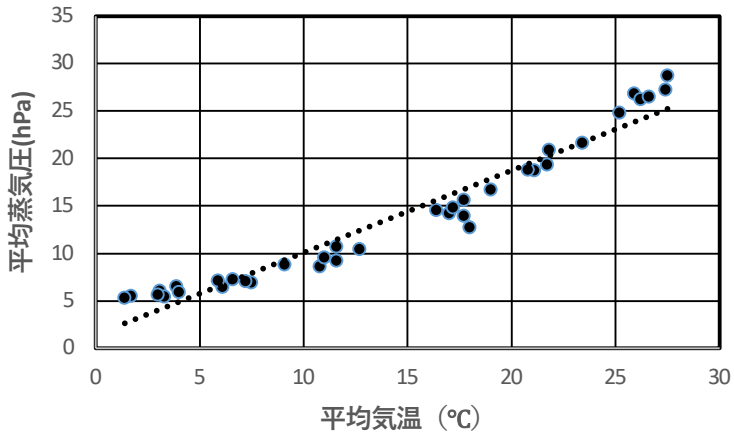
NIIGATA UNIVERSITY

- **分類・回帰問題**などの入出力の**パターン学習方法**
- 現AIのコア技術．ここ数年は**生成系**が流行っている．
- 特にNeural Network系はこの10年で非常に高精度に．



機械学習（入出力パターンの学習）

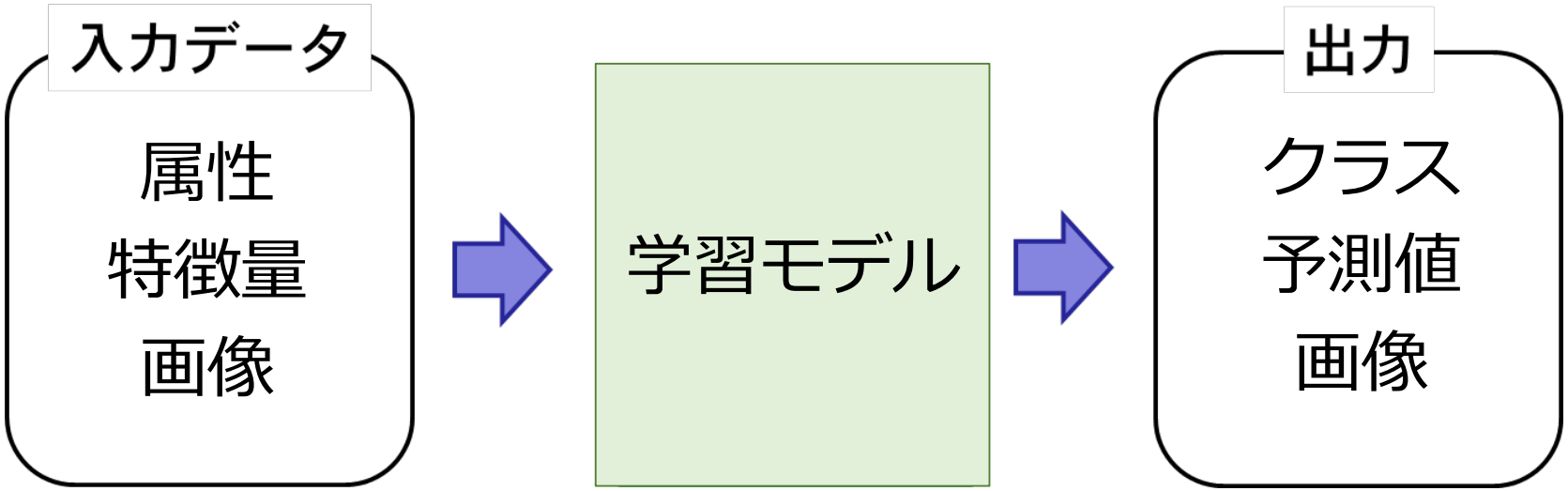




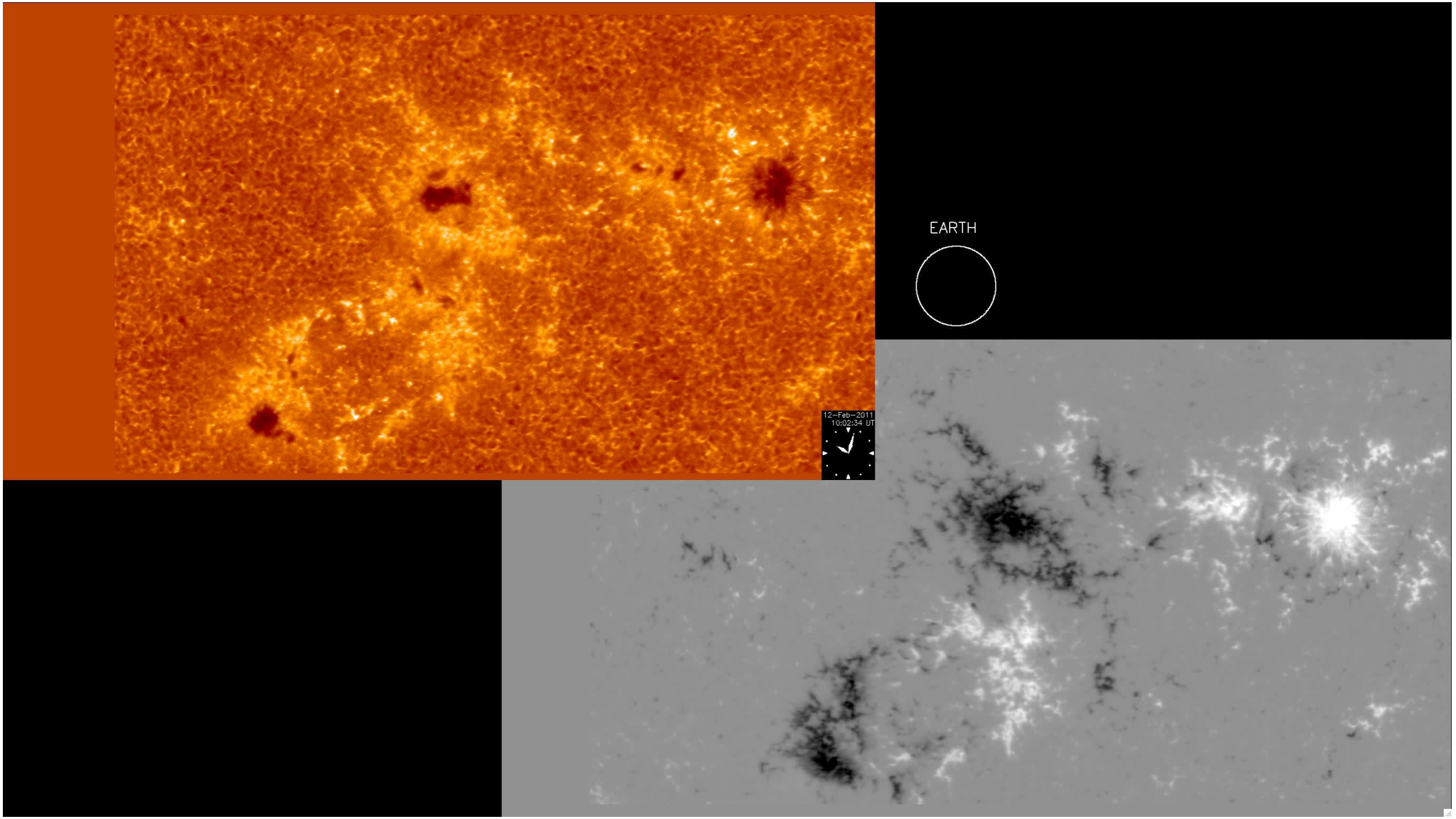
最小二乗法（回帰分析）も
機械学習の１種

機械学習モデルの作り方

1. **問題設定**（入力，出力 etc.）の決定
2. **アルゴリズム**（RF, SVM, 深層学習, etc.）の決定
3. モデルアーキテクチャの決定
4. **学習方法(損失関数，最適化方法)**の決定と実行
5. **汎化性能**の評価



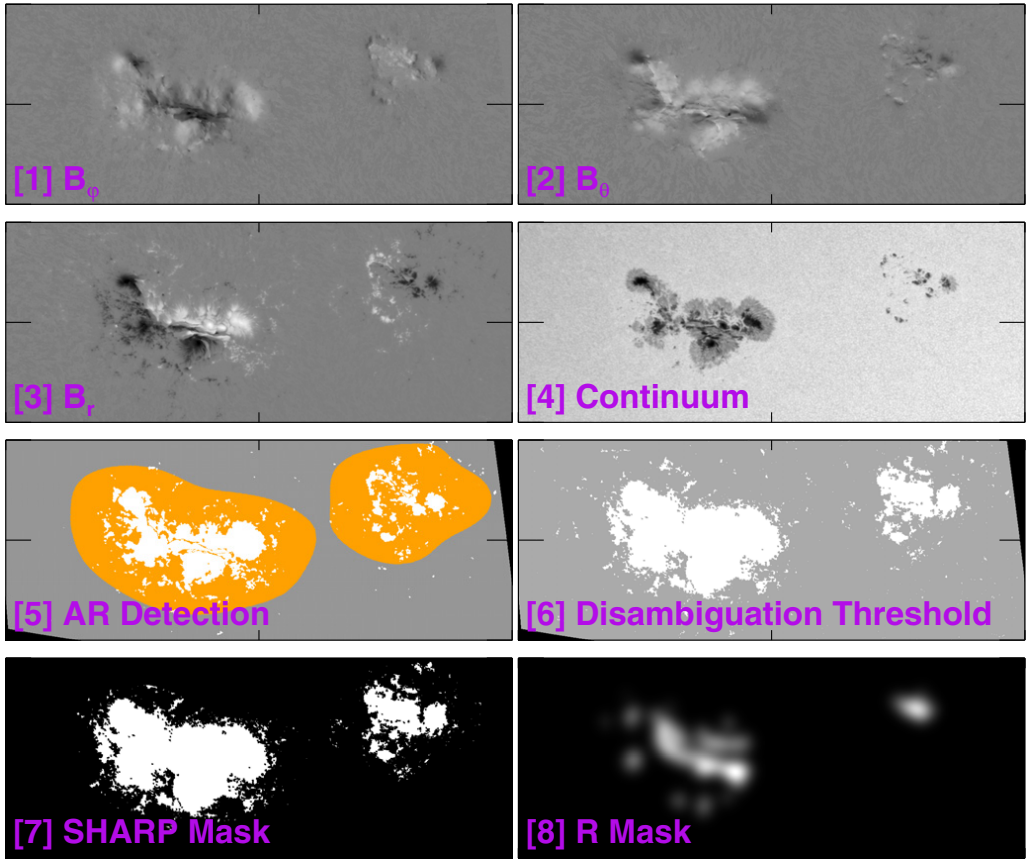
研究例１　太陽フレアの予測



機械学習によるフレア予測 (Bobra+, 2015)

NIIGATA UNIVERSITY

- 約150万枚の活動領域画像から予測モデルを構築
- 人を超えたTSS=0.76を達成　　cf) TSS～0.5？@研究者



$$TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$$

Prediction	Flare	No flare
Flare	TP	FP
No flare	FN	TN

Flare　　No flare

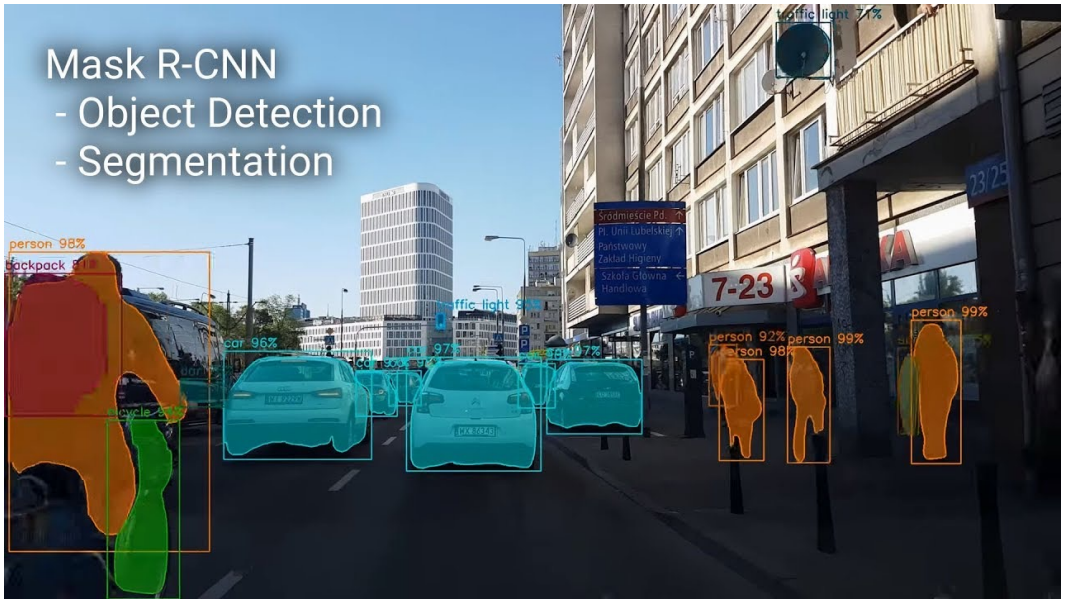
Ground truth

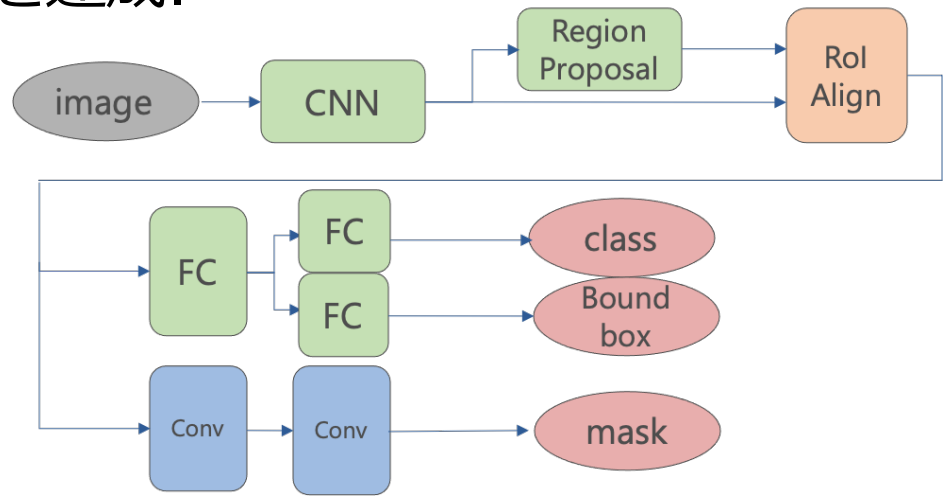
研究例2 全球磁場画像からのフレア予測

NIIGATA UNIVERSITY

Mask R-CNN（He et al, 2017）

・**マスク学習ブランチ**を持つ，物体検出モデル.
・３つの損失関数(BB, MSK, CLS)を最適化する.
・COCO分類問題において最高スコアを達成.



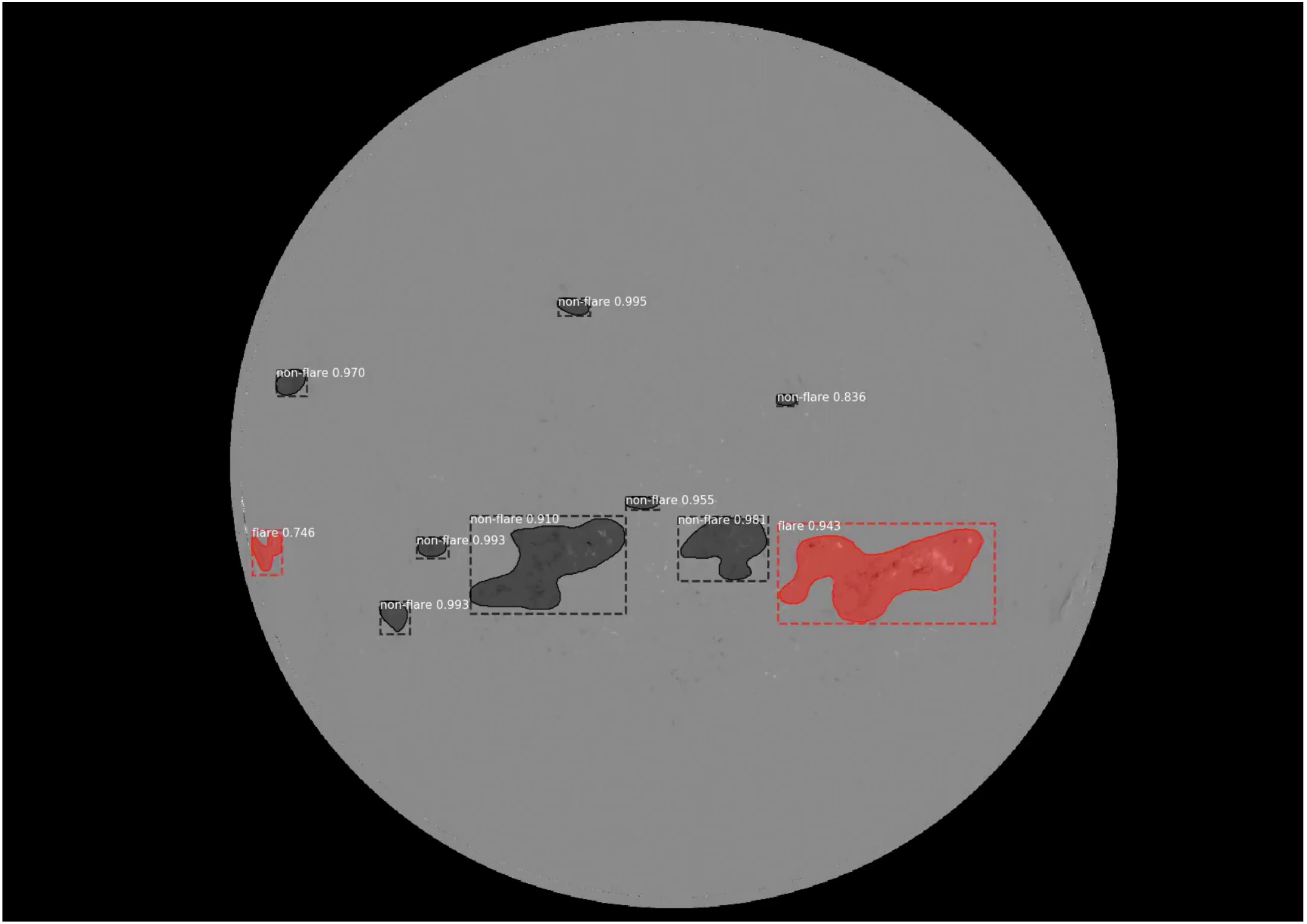


$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$$

$$\mathcal{L}(\{p_i\},\{t_i\}) = \frac{1}{N_{\text{cls}}}\sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{box}}}\sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*)$$

$$\mathcal{L}_{\text{cls}}(p_i, p_i^*) = -p_i^* \log p_i - (1 - p_i^*)\log(1 - p_i)$$

$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2}\sum_{1\le i,j\le m}\left[y_{ij}\log\hat{y}_{ij}^k + (1 - y_{ij})\log(1 - \hat{y}_{ij}^k)\right]$$


non-flare 0.995
non-flare 0.970
non-flare 0.836
non-flare 0.955
non-flare 0.910
non-flare 0.981
flare 0.943
flare 0.746
non-flare 0.993
non-flare 0.993

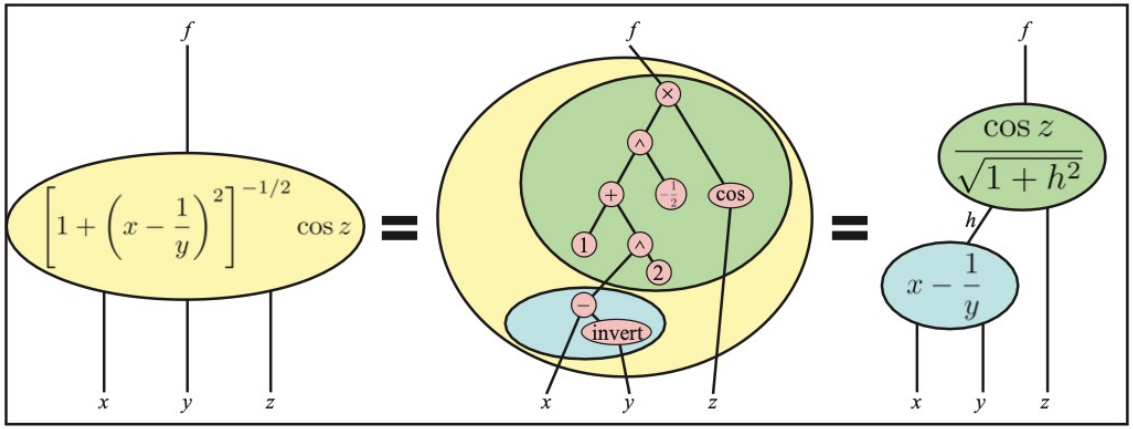
研究例３ 関数同定による物理法則の発見

ニューラルネットワークによる高精度なデータ内挿･外挿を用いた**データの対称性・分割性から**効率的な関数の同定を行う

ファインマン物理学の100個の方程式を**2時間以内に全て解いた**

研究目的

AI-Feynmanを拡張して
微分方程式に対応する



Feynman eq.	Equation
I.6.20a	$f = e^{-\theta^2/2}/\sqrt{2\pi}$
I.6.20	$f = e^{-\frac{\theta^2}{2\sigma^2}}/\sqrt{2\pi\sigma^2}$
I.6.20b	$f = e^{-\frac{(\theta-\theta_1)^2}{2\sigma^2}}/\sqrt{2\pi\sigma^2}$
I.8.14	$d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$
I.9.18	$F = \frac{Gm_1m_2}{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$
I.10.7	$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$
I.11.19	$A = x_1y_1 + x_2y_2 + x_3y_3$
I.12.1	$F = \mu N_n$

AI-Feynmanが解いたファインマン物理学の式の一例

右下表 : [1] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, Vol 6, Issue 16, 2020.
右上図 : [2] Silviu-Marian Udrescu and Max Tegmark. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. arXiv:2006.10782[cs.LG].

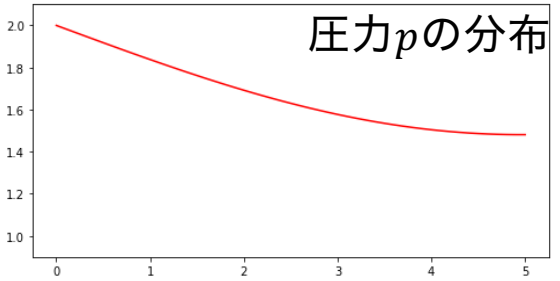
研究例３ 関数同定による物理法則の発見

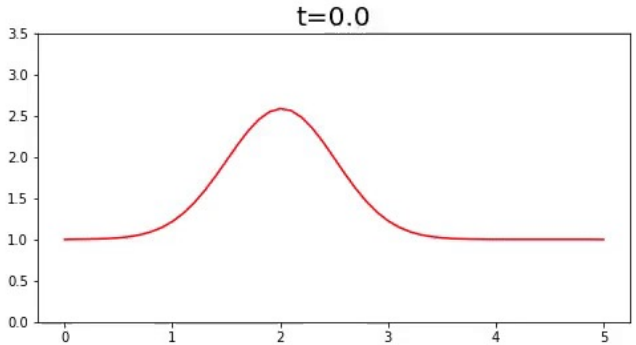
- 空間を0~5に設定し，50分割：$dx = 0.1$
- 時間刻みは0.005(s)：$dt = 0.005$
- 拡散係数：$\nu = 0.3$
- 流体の密度：$\rho = 0.4$
- **説明変数は$u, p, \frac{\partial u}{\partial x}, \frac{\partial p}{\partial x}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 p}{\partial x^2}$の6つ**

ナビエ・ストークス方程式

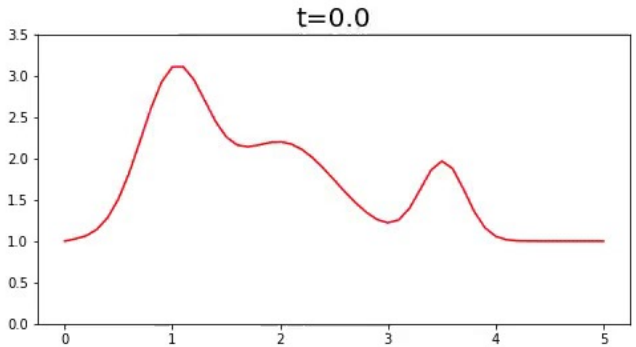
$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\frac{\partial^2 u}{\partial x^2}$$

時間変化項　移流項　圧力項　拡散項

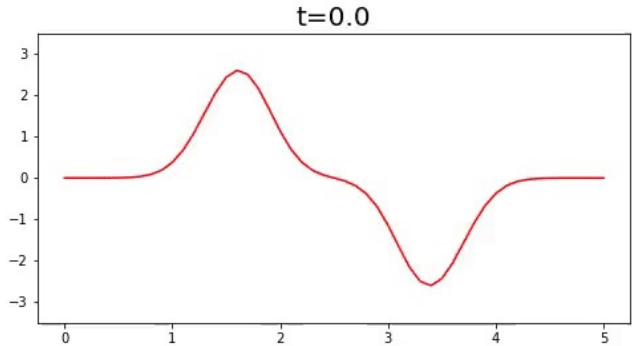




分布(1)



分布(2)



分布(3)

研究例３ 関数同定による物理法則の発見

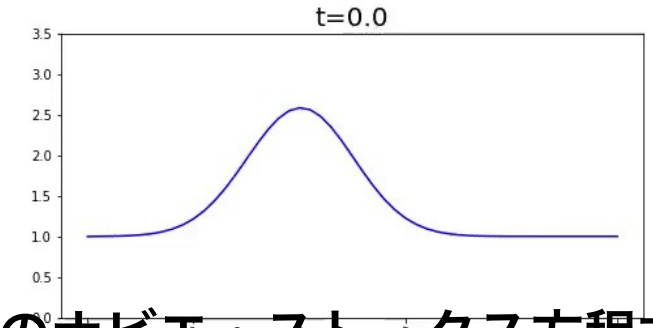
分布	移流方程式	バーガース方程式 $(\nu = 0.1)$	ナビエ・ストークス方程式 $(\nu = 0.3, \rho = 0.4)$
(1)	○ $-u\frac{\partial u}{\partial x}$	○ $-u\frac{\partial u}{\partial x} + 0.1\frac{\partial^2 u}{\partial x^2}$	△ $-0.975u\frac{\partial u}{\partial x} + 0.318\frac{\partial^2 u}{\partial x^2} + 0.310$
(2)	○ $-u\frac{\partial u}{\partial x}$	○ $-u\frac{\partial u}{\partial x} + 0.1\frac{\partial^2 u}{\partial x^2}$	△ $-1.001u\frac{\partial u}{\partial x} + 0.312\frac{\partial^2 u}{\partial x^2} + 0.312p - 1.001u\frac{\partial p}{\partial x} - 0.450$
(3)	○ $-u\frac{\partial u}{\partial x}$	○ $-u\frac{\partial u}{\partial x} + 0.1\frac{\partial^2 u}{\partial x^2}$	△ $-u\frac{\partial u}{\partial x} + 0.318\frac{\partial^2 u}{\partial x^2} + 0.382$
実行時間 (平均)	1時間9分	1時間7分	2時間30分

正解の式：$-u\frac{\partial u}{\partial x} + 0.3\frac{\partial^2 u}{\partial x^2} - 2.5\frac{\partial p}{\partial x}$

移流項　拡散項　圧力項

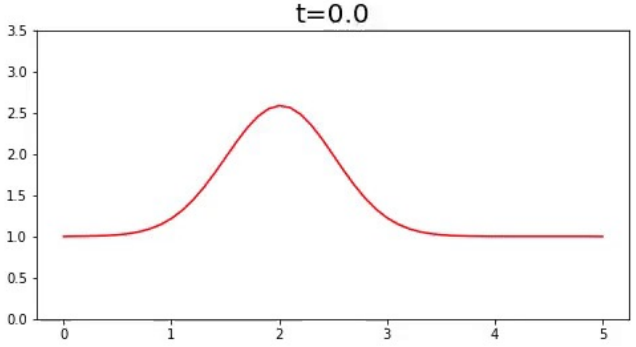
研究例３ 関数同定による物理法則の発見

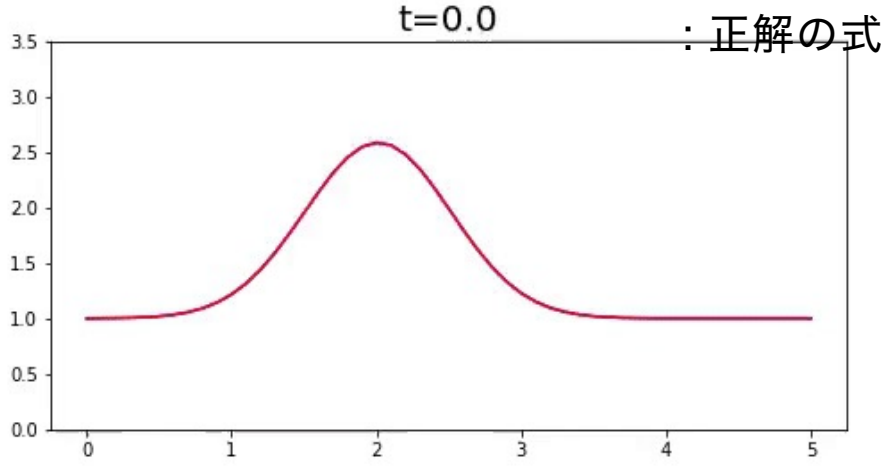
AI-Feynman : $\frac{\partial u}{\partial t} = -0.975u\frac{\partial u}{\partial x} + 0.318\frac{\partial^2 u}{\partial x^2} + 0.310$



正解のナビエ・ストークス方程式 :

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - 2.5\frac{\partial p}{\partial x} + 0.3\frac{\partial^2 u}{\partial x^2}$$



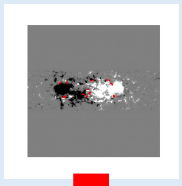


⇒ 式の形は異なるが、**データを**
ほぼ正確に記述できる微分方程式を同定

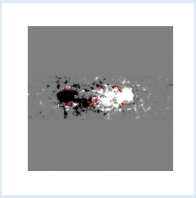
新しい物理モデリング方法の提案

深層学習による非線形現象モデリングはやはり強力．さらに，**データからの関数同定**も現実的になってきた．以下のモデリングが可能？

Step1 深層学習による予測モデル構築



DLモデル



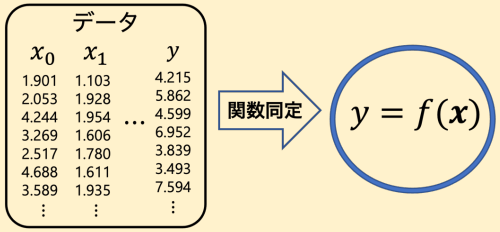
できるだけ高精度な予測モデルを作る

数値シミュレーション
- なぜだかできるものの具体的に法則式を知りたいもの？

観測データ
- 数値シミュレーションでは再現できないもの？

ブラックボックスだが
(高精度の）データ生成器

Step2 AI Feynman(など)による関数同定

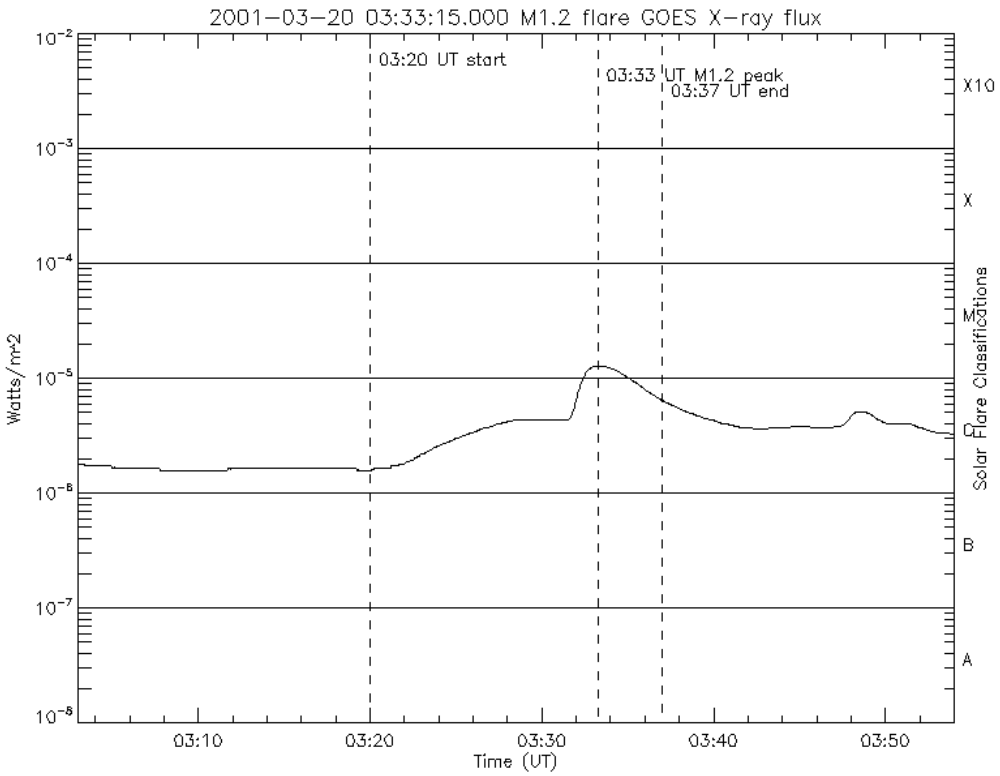


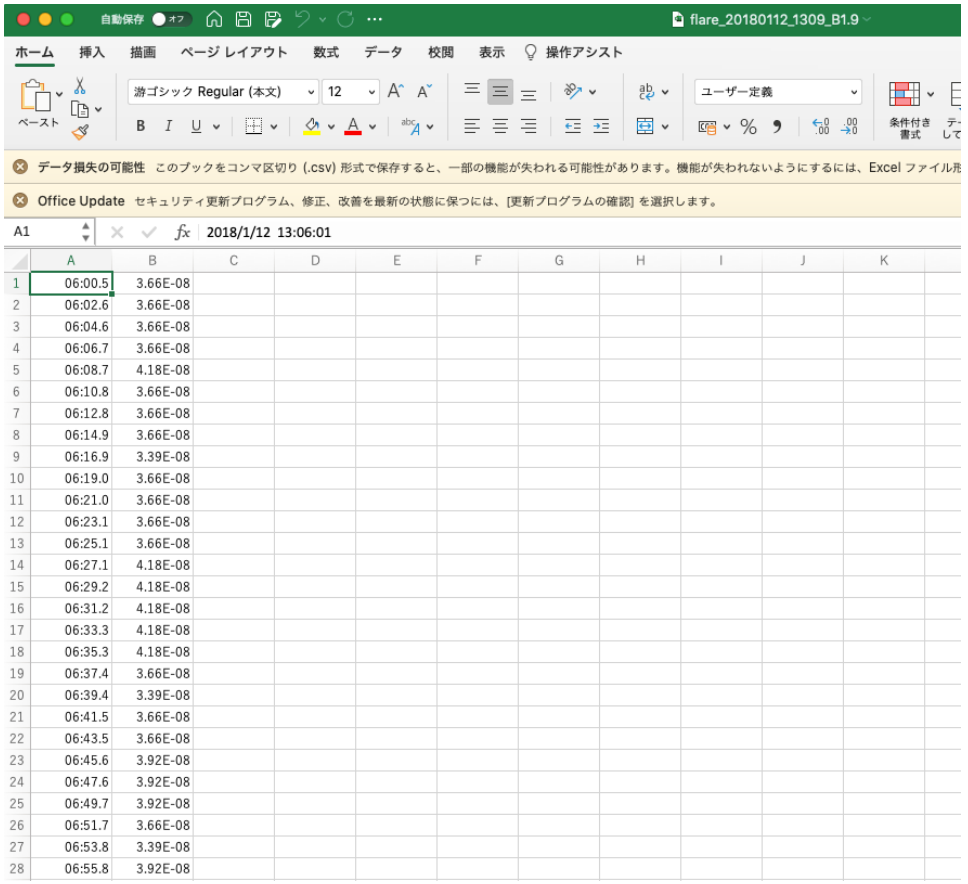
関数系は簡略化
- 知っている項は引いておく

データはできるだけ少なく
- 情報論的エントロピーなどに基づいたサンプリング

機械学習の実習

GOES X-ray Fluxデータからの**フレアクラス分類問題**
→ 何をするかノーヒントで，**データ最大値を読み取り分類する**





	A	B
1	06:00.5	3.66E-08
2	06:02.6	3.66E-08
3	06:04.6	3.66E-08
4	06:06.7	3.66E-08
5	06:08.7	4.18E-08
6	06:10.8	3.66E-08
7	06:12.8	3.66E-08
8	06:14.9	3.66E-08
9	06:16.9	3.39E-08
10	06:19.0	3.66E-08
11	06:21.0	3.66E-08
12	06:23.1	3.66E-08
13	06:25.1	3.66E-08
14	06:27.1	4.18E-08
15	06:29.2	4.18E-08
16	06:31.2	4.18E-08
17	06:33.3	4.18E-08
18	06:35.3	4.18E-08
19	06:37.4	3.66E-08
20	06:39.4	3.39E-08
21	06:41.5	3.66E-08
22	06:43.5	3.66E-08
23	06:45.6	3.92E-08
24	06:47.6	3.92E-08
25	06:49.7	3.92E-08
26	06:51.7	3.66E-08
27	06:53.8	3.39E-08
28	06:55.8	3.92E-08