「宇宙科学から統計数理を経てビジネスへ 」
-- データサイエンスを軸として --

Datum Studio(株)

森井幹雄

2022.02.18 (Fri), 13:15-14:00, ISAS/JAXA

自己紹介

- 森井幹雄： Data Scientist @ DATUM STUDIO株式会社
- 東京工業大学、物理学科(1995 – 2003)
  - 修士課程: Super-Kamiokande実験のデータ解析
  - 博士課程: X線天文学(河合研究室): マグネターの観測
- 博士取得後、観測天文学のポスドク(2004 – 2014)
  - JAXA(筑波宇宙センター), 立教大学, 東京工業大学, 理化学研究所
  - 全天X線監視装置(MAXI)プロジェクトに参加
    - 装置開発、較正試験、ソフトウェア開発、運用、観測データ解析
  - X線突発天体の研究
- 統計数理研究所：ポスドク、特任助教(2015 – 2019)
  - Subaru望遠鏡のHyper Sprime-Cam で得られるビッグデータの解析
  - 統計的手法を天文学データ解析に応用する研究。
- 現職：IT企業@虎ノ門ヒルズ（ドラゴン桜のロケで使用; 林遣都社長）
  - Neural Network の論文実装
  - FAB for HMEの論文実装
  - 最短経路問題を用いた製造工程の最適化

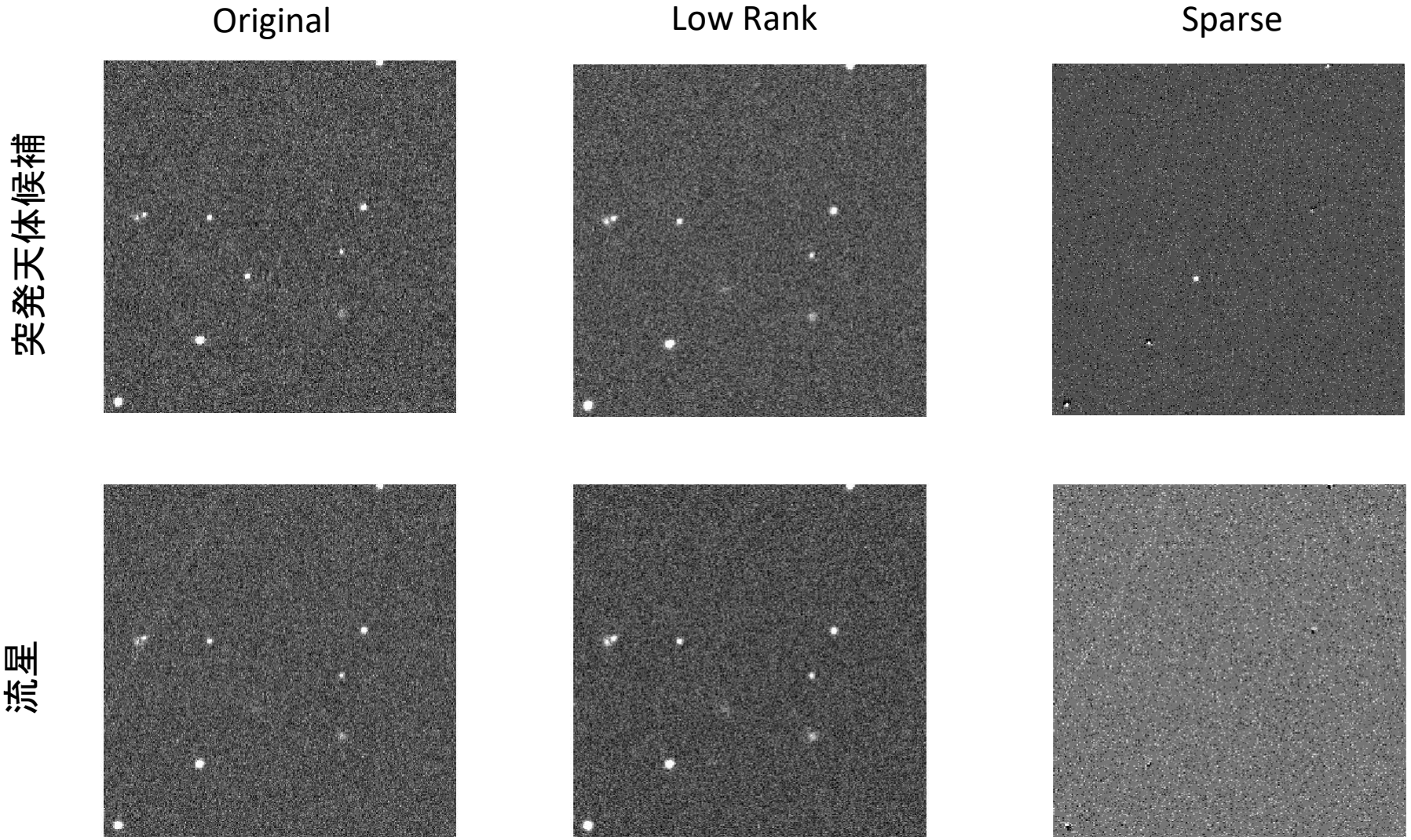
講演に至る経緯

- 座長の海老沢さんと森井はX線天文学の研究者という共通点があり、MAXIプロジェクトで共同研究を行いました。
- 統計数理研究所の研究員時代に、「宇宙科学情報解析シンポジウム」で2回講演をしています。
  - 2016年「突発天体探索の手法について」
  - 2019年「Angular Resolution Booster を用いたX線望遠鏡のイメージ再構成法」
- 2019年にIT企業に就職し、天文学の研究からは離れましたが、私のような経歴の人は珍しいので講演して欲しいということで、講演をお引き受けいたしました。

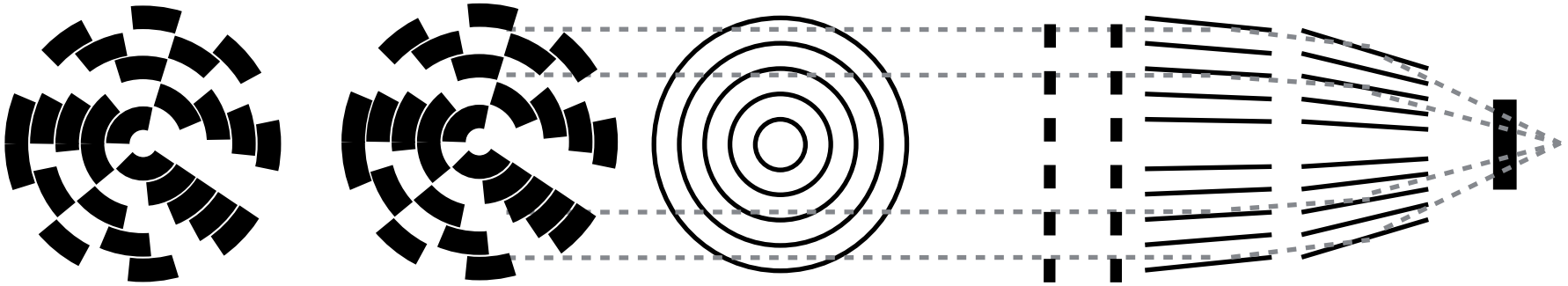
2016年の講演:「突発天体探索の手法について」

低ランク行列分解の応用

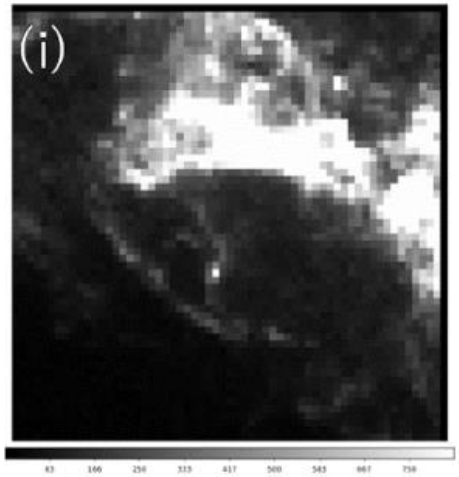
Morii et al. (2017), ApJ, 835, 1, “Data Compression for the Tomo-e Gozen Using Low-rank Matrix Approximation”



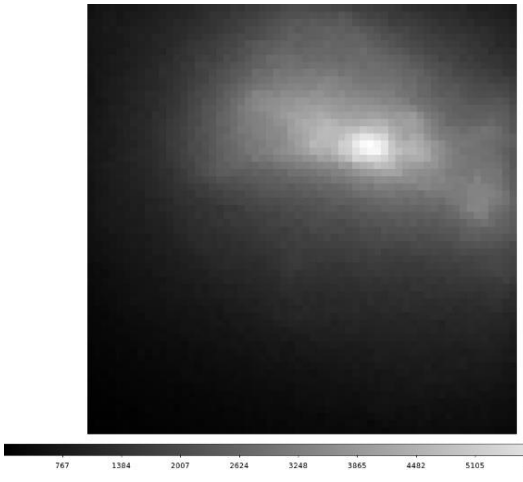
2019年「Angular Resolution Booster を用いたX線望遠鏡のイメージ再構成法」



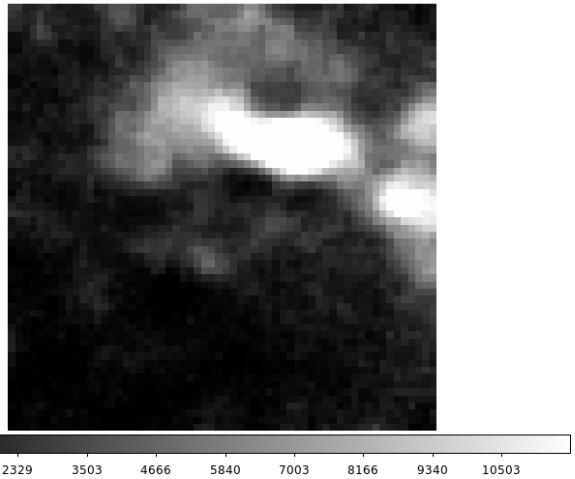
正解のイメージ



Booster なし



Booster あり



- Morii, Ikeda & Maeda (2019), PASJ, 71, 24, “An image reconstruction method for an X-ray telescope system with an angular resolution booster”
- Maeda et al. (2019), PASJ, 71, 97, “Concept for an X-ray telescope system with an angular resolution booster”

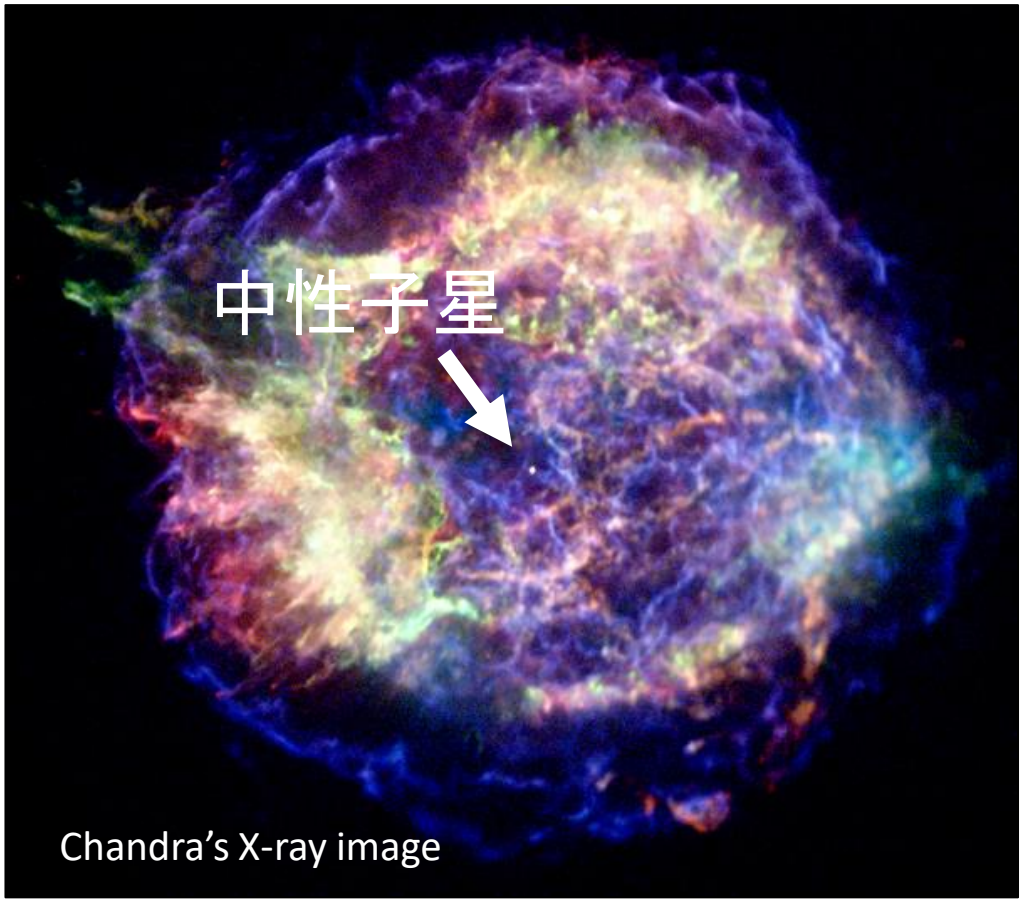
内容

- 統計数理研究所時代に実施した超新星残骸のデータ解析の紹介
  - 勝田哲さん（埼玉大学）との共同研究
  - 佐藤寿紀さん（立教大学）との共同研究
- 弊社にて行ったデータ科学の応用例の紹介
  - EUNNの実装
  - FAB for HMEの実装
  - 最短経路問題を用いた製造工程の最適化
  - ベイズ最適化
- ポスドクから民間企業へ移って分かったこと

超新星残骸 (SNRs)

- 太陽の約10倍以上の質量を持つ大質量星は寿命の最後に超新星爆発を起こす。
- 周囲に星の外層をばらまき、中心に中性子星(NS)やブラックホールを生成する。
- 中性子星は爆発時にkickされ高速で飛び出すことが観測されている。
- 中性子星のkickメカニズムとして2つの説が提唱されている。

超新星残骸 Cas A のイメージ

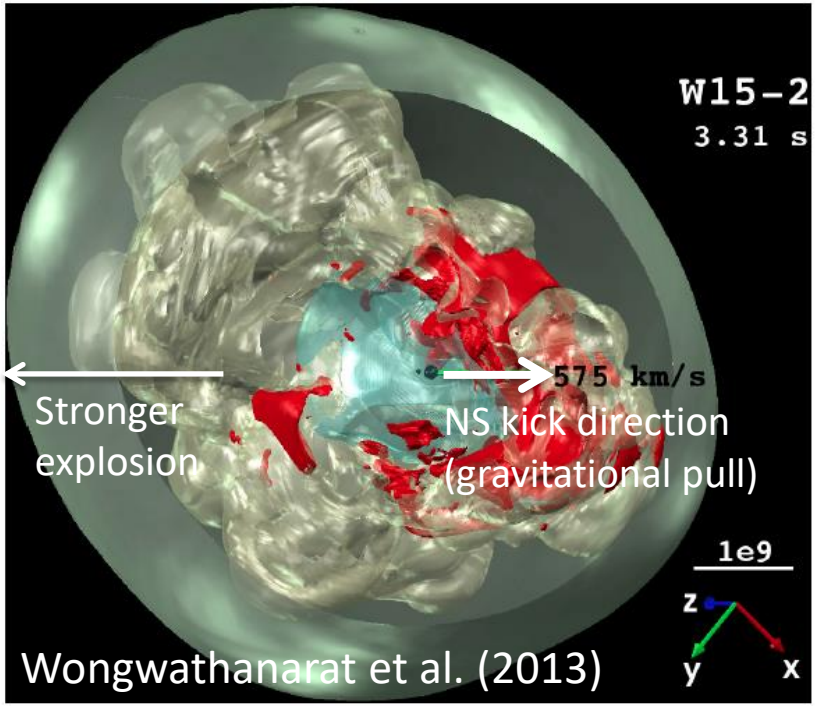


勝田哲さん（埼玉大学）との共同研究

中性子星のキックメカニズム

Hydrodynamic kick

Explosion asymmetry kicks NS.

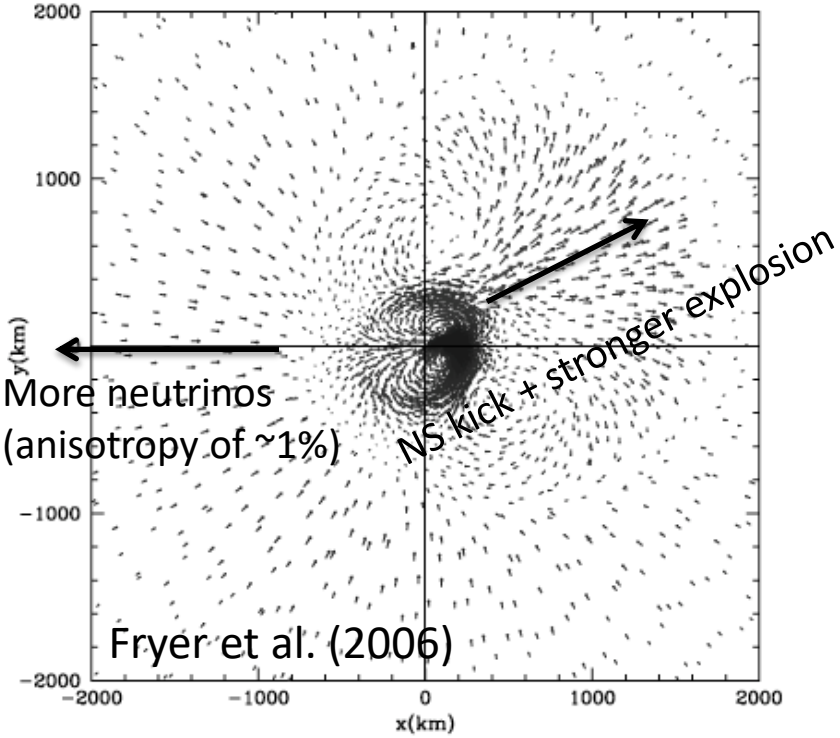


Ejecta ⇔ NS

残骸重心と中性子星位置は、
爆発中心に対して対称になる。

Neutrino-induced kick

Anisotropic neutrino emission kicks NS.

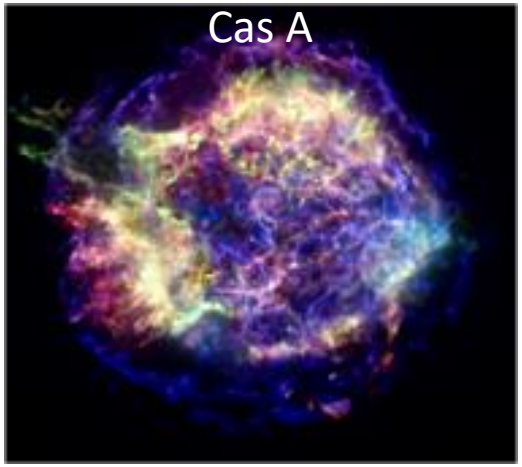


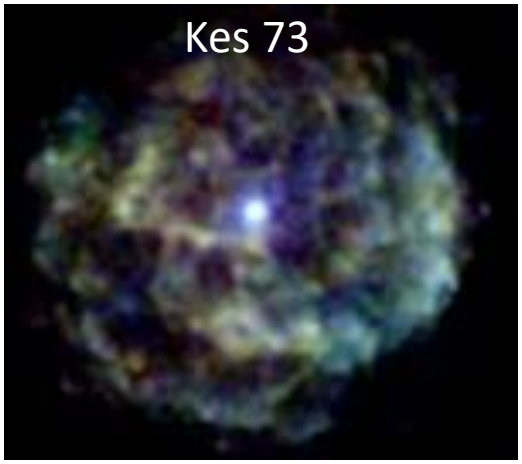
Ejecta + NS ⇔ Neutrino

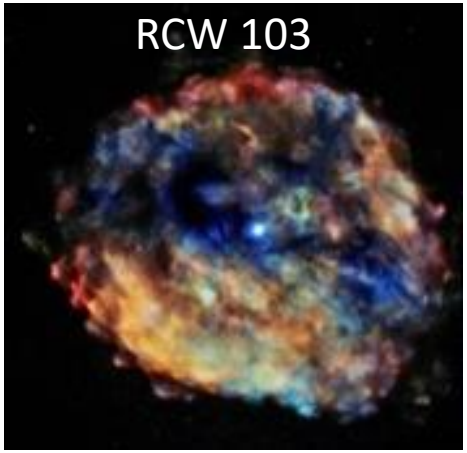
残骸重心と中性子星位置は、
爆発中心に対して非対称になる。

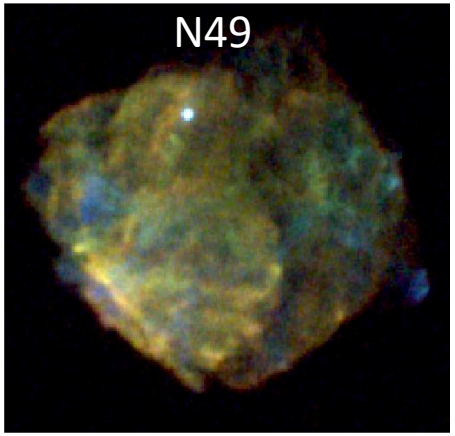
勝田哲さん（埼玉大学）との共同研究

対象天体









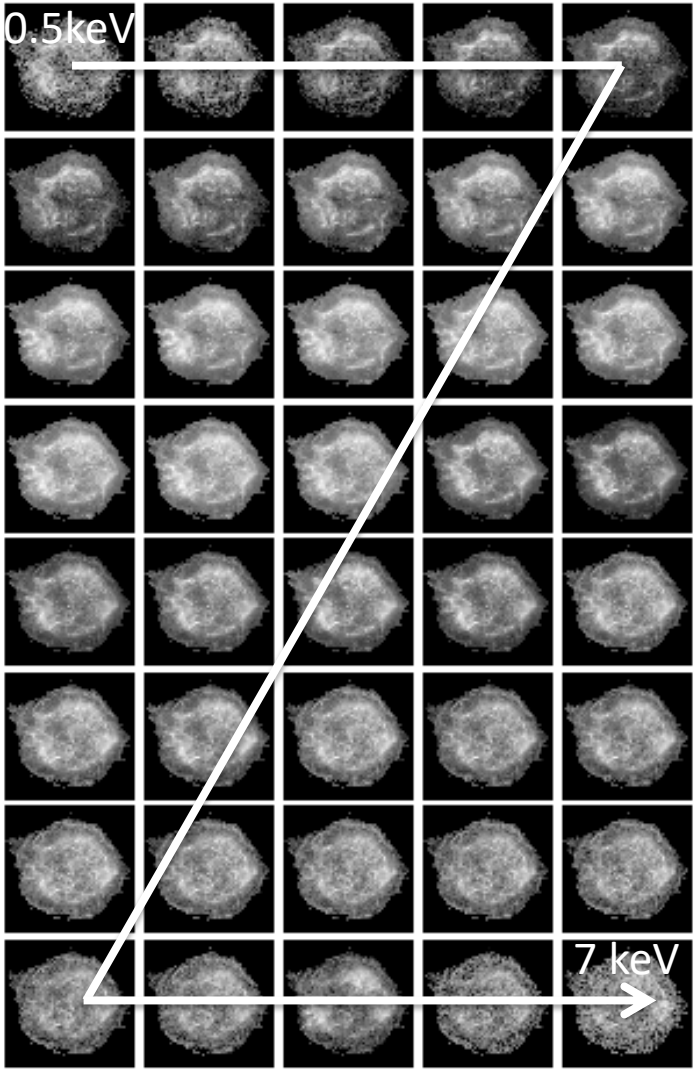
中性子星と爆発中心の位置が分かっている6天体について調査

放出物の重心位置を求めたい。

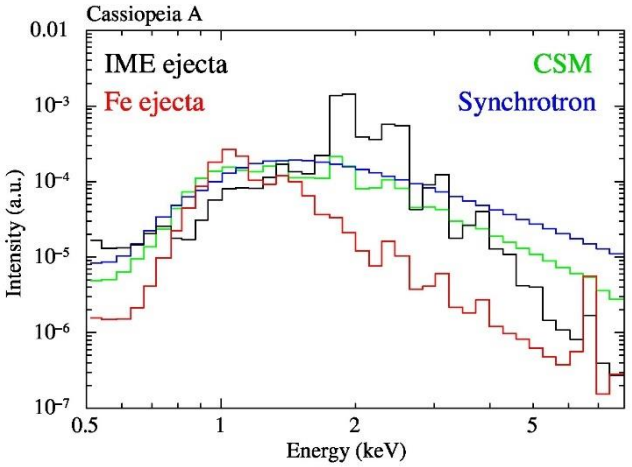
勝田哲さん（埼玉大学）との共同研究

放出物(Ejecta)成分のイメージを作成する
Image Decomposition

1. Narrow band images



2. Templates of individual spectral components (e.g., IME-rich ejecta, CSM)



Input 1

Input 2

Spectrum from each pixel

各ピクセル毎に、各スペクトル成分の量を求める。Ejecta成分だけのイメージを作成する。

勝田哲さん（埼玉大学）との共同研究

Fitting method

- 通常の方法:
  - Xspec:　Maximum Likelihood + Local optimization（MINUIT etc.）
  - 非常に手間と時間がかかる。
- 我々の方法:
  - χ^2_γ Statistics + Quadratic Programming

勝田哲さん（埼玉大学）との共同研究

χ^2_γ Statistics + Quadratic Programming

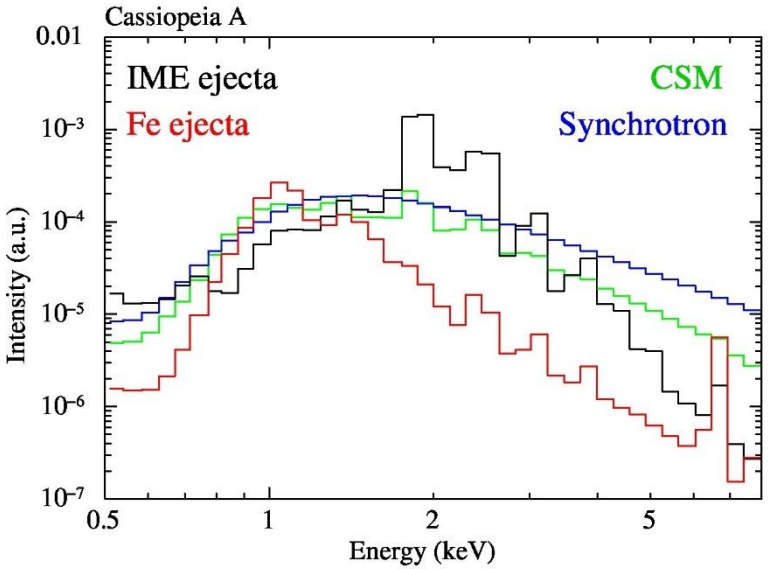
- Numerical Recipes in C で紹介されていた近似法を用いた。
- Approximation by quadratic form proposed by K. J. Mighell 1999, ApJ, 518, 380.
- フラックス >= 0 の制約。
- 低カウント数のピクセルで問題が生じる。
- Quadratic Programmingを用いることで解決。

$$\chi^2_\gamma = \sum_{i=1}^{N} \frac{(n_i + \min(n_i, 1) - m_i)^2}{n_i + 1}.$$

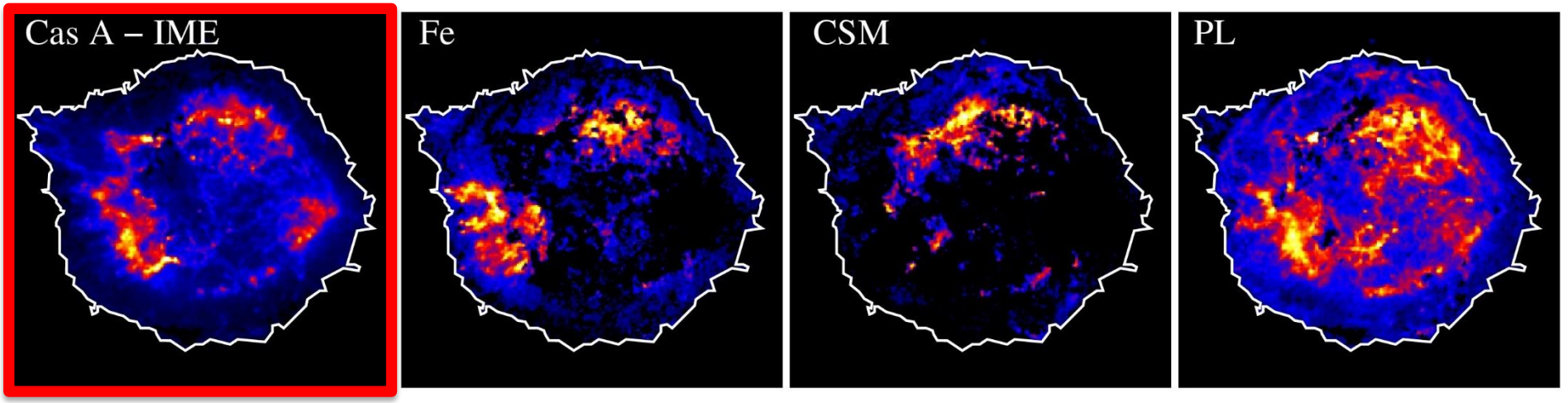
$$\begin{aligned} \text{Minimize} \quad &: \quad \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x}, \\ \text{subject to} \quad &: \quad A\mathbf{x} = \mathbf{b}, \\ &\quad\quad\; C\mathbf{x} \leq \mathbf{d}. \end{aligned}$$

勝田哲さん（埼玉大学）との共同研究

Image Decomposition for Cas A

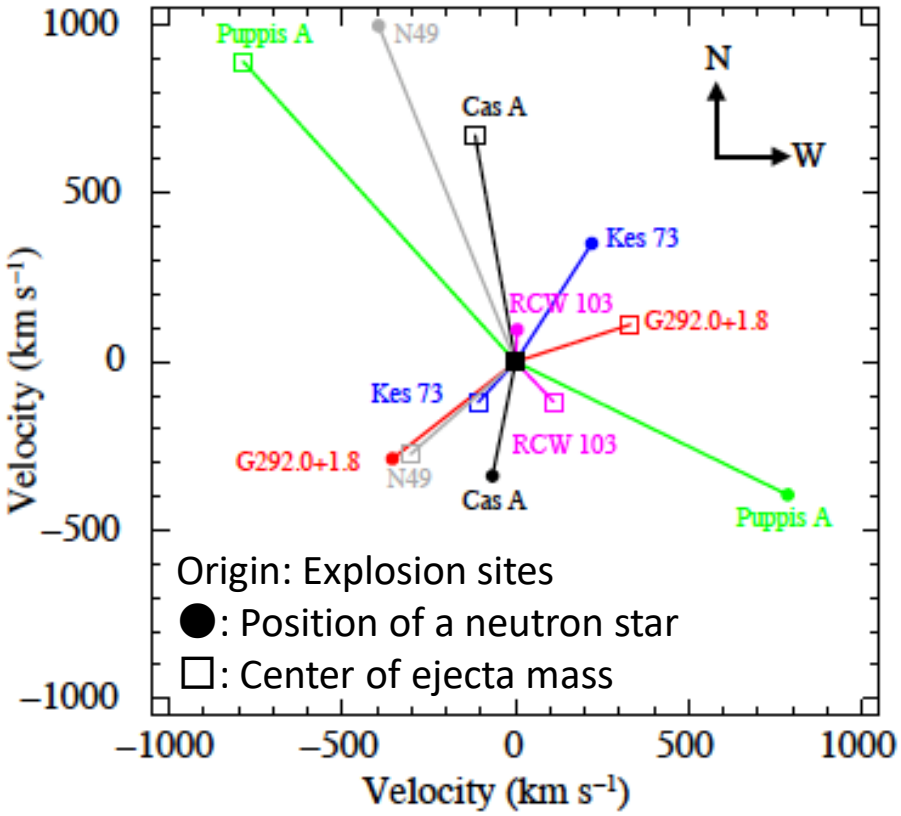


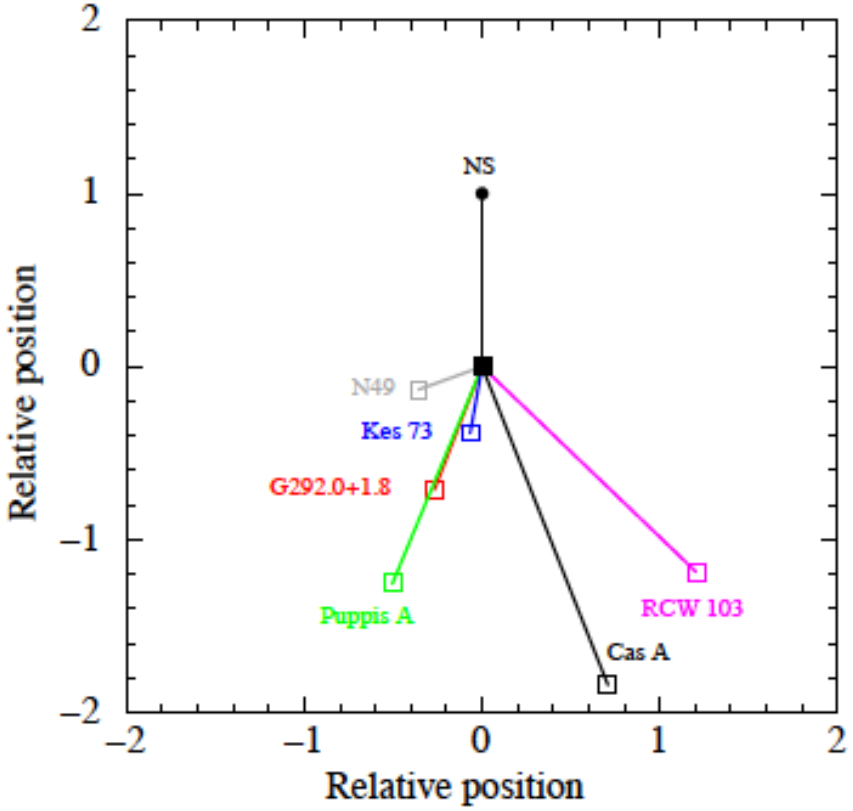
IME distribution



勝田哲さん（埼玉大学）との共同研究

爆発中心、中性子星、Ejecta重心位置のプロット





明らかに、hydrodynamic kick モデルを支持する結果を得た。

Katsuda et al. (2018), ApJ, 856, 18, “Intermediate-mass Elements in Young Supernova Remnants Reveal Neutron Star Kicks by Asymmetric Explosions”

勝田哲さん（埼玉大学）との共同研究

アイデアの源泉

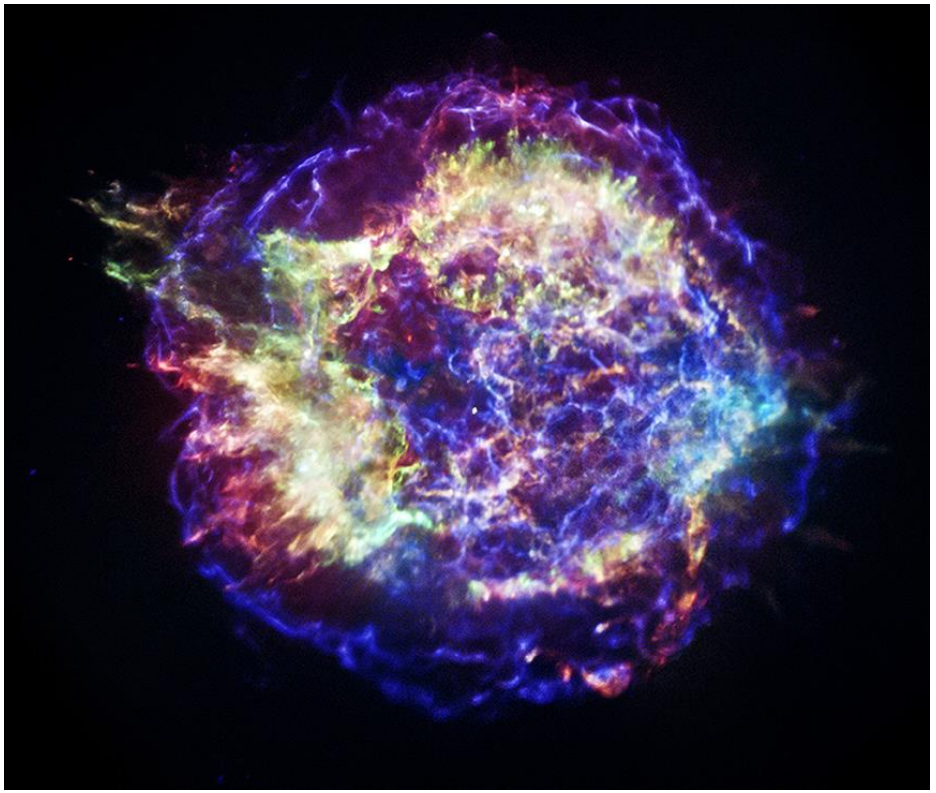
- MAXIのデータ解析で天体の光度をイメージフィットによって求める課題に取り組んでいたときに考案しました。
- Numerical Recipes in C を読んで、低カウントの場合に使えるχ^2 の補正式として、χ^2_ν を用いる方法を知った。しかし、それだけでは、フラックスが負になる問題は解決しない。
- Cern/ROOT のライブラリを調べていたら、quadratic programming が紹介されており、それを用いれば、フラックス非負の制約の元で解けることに気付いた。
- MAXIのイメージフィットのプログラムに組み込んでこの方法を試したことがあったが、既に従来の方法でフィットするプログラムが完成していたので必要なかった。
- Tomo-e Gozenの研究会で、勝田さん達と超新星残骸の解析について飲みながら議論しているときに、MAXIでは活用できなかったこの方法が応用できると気づいた。

勝田哲さん（埼玉大学）との共同研究

超新星残骸の逆行衝撃波の抽出

- 約350年前に発生した超新星爆発の残骸。
- 銀河系内の若い超新星残骸の中では最もよく観測され研究が進んでいる天体。
- Chandra　X線衛星による10年以上に渡る観測により、Cas Aが膨張する様子が動画で得られている。

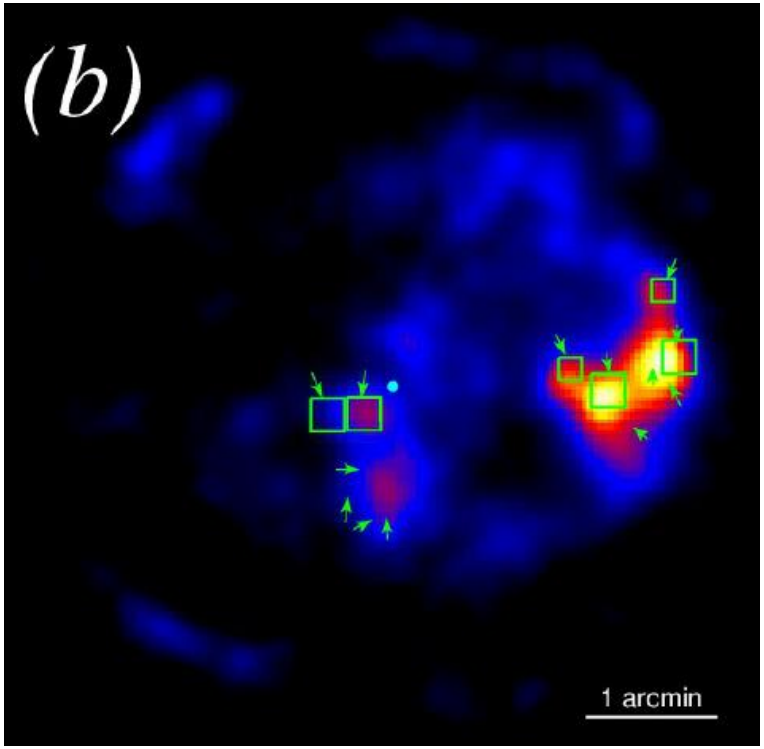
超新星残骸 Cas A のイメージ



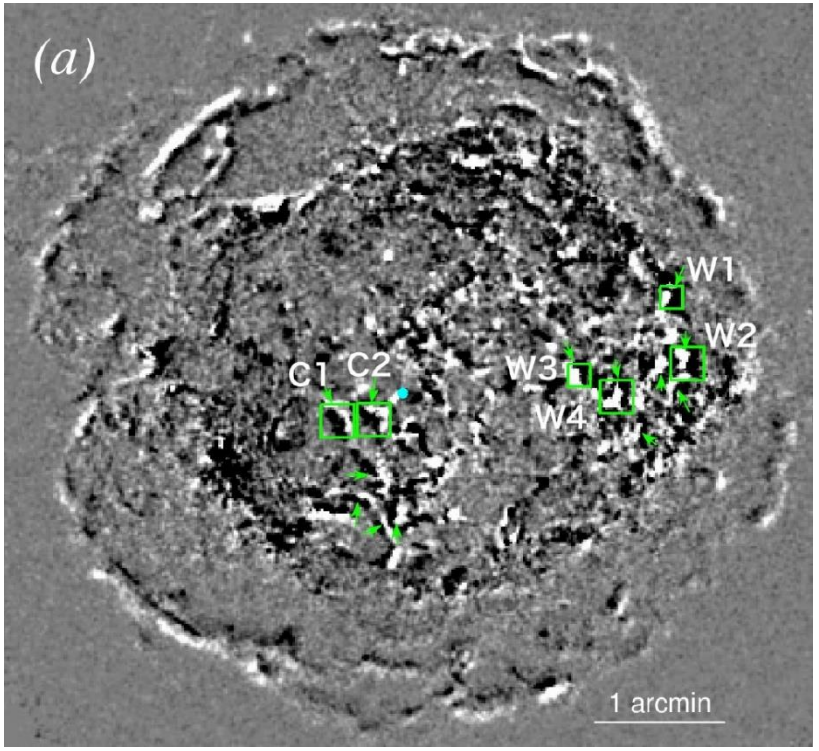
- 超新星残骸の波面は高エネルギー宇宙線の加速源と考えられている。

佐藤寿紀さん（立教大学）との共同研究

Cas A のイメージ



NuSTAR 衛星で取得した硬X線放射
(15-40 keV)



2000年と2014年にChandra衛星により
取得したイメージの差分 (4.2 – 6 keV)

逆行している領域と、硬X線放射領域が対応しているようだが本当か？

佐藤寿紀さん（立教大学）との共同研究

Optical flow

移動物体の検出や、その動作の解析などに用いられる方法

http://cs.brown.edu/courses/csci1290/2011/results/final/psastras/images/sequence0/save_0.png

佐藤寿紀さん（立教大学）との共同研究

Farneback's Alogorithm

イメージが局所的に2次式で近似できるとする。

$$f_1(\mathbf{x}) = \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + \mathbf{b}_1^T \mathbf{x} + c_1$$

これを、d だけ平行移動すると、

$$\begin{aligned} f_2(\mathbf{x}) = f_1(\mathbf{x} - \mathbf{d}) &= (\mathbf{x} - \mathbf{d})^T \mathbf{A}_1 (\mathbf{x} - \mathbf{d}) + \mathbf{b}_1^T (\mathbf{x} - \mathbf{d}) + c_1 \\ &= \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d})^T \mathbf{x} + \mathbf{d}^T \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^T \mathbf{d} + c_1 \\ &= \mathbf{x}^T \mathbf{A}_2 \mathbf{x} + \mathbf{b}_2^T \mathbf{x} + c_2. \end{aligned}$$

となり、これも2次式。

係数を比較して、

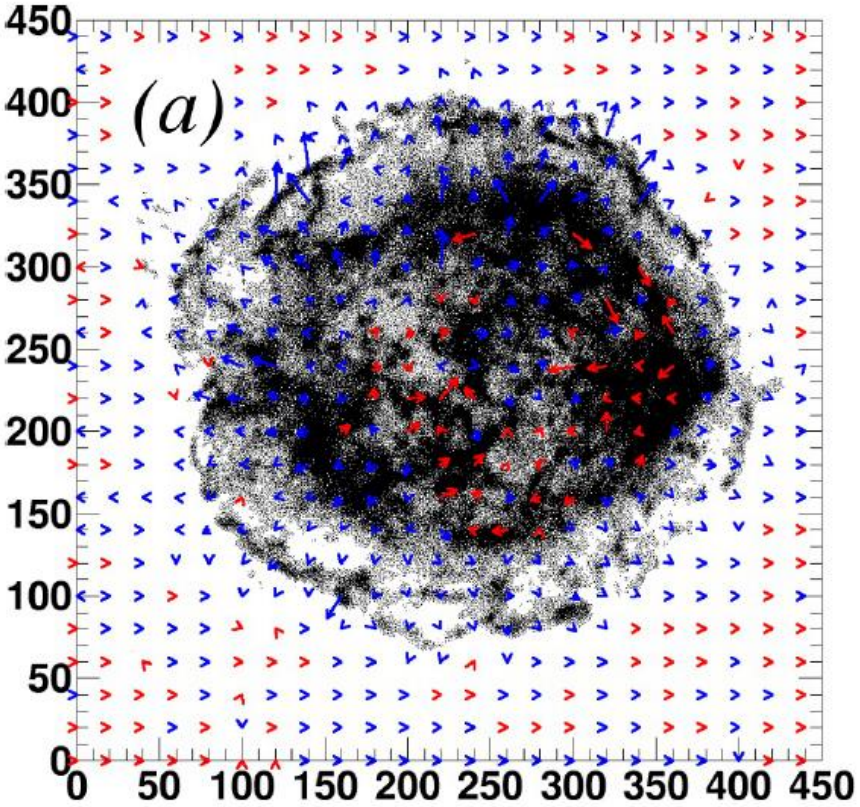
$$\begin{aligned} \mathbf{A}_2 &= \mathbf{A}_1, \\ \mathbf{b}_2 &= \mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d}, \\ c_2 &= \mathbf{d}^T \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^T \mathbf{d} + c_1. \end{aligned}$$

A_1 が非特異であれば、

$$\mathbf{d} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1).$$

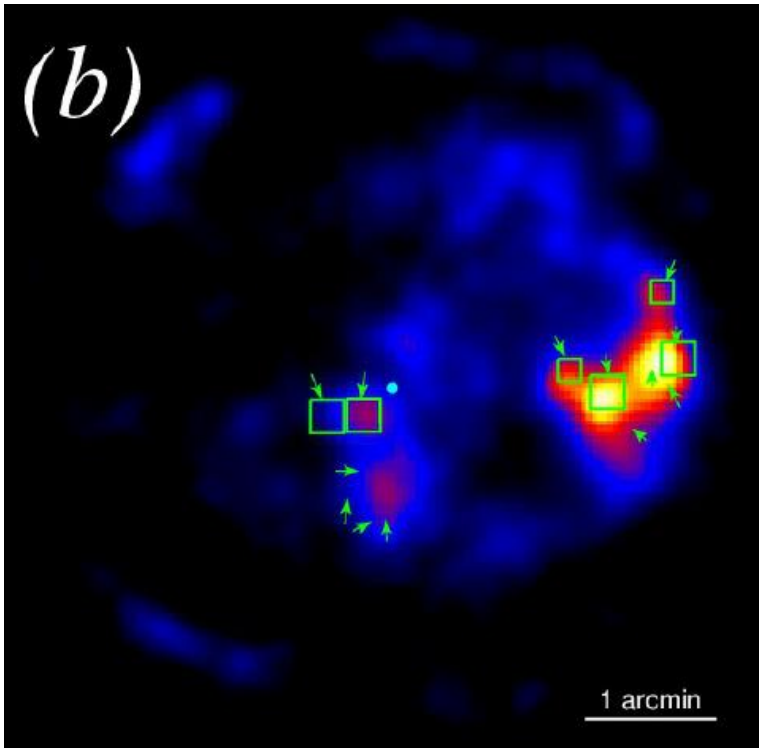
で移動ベクトルdが求まる。

Optical flow による速度場と硬X線放射マップの比較



Chandra衛星 (2000年→ 2014年)

Synchrotron 放射 (4.1 – 6.3 keV)



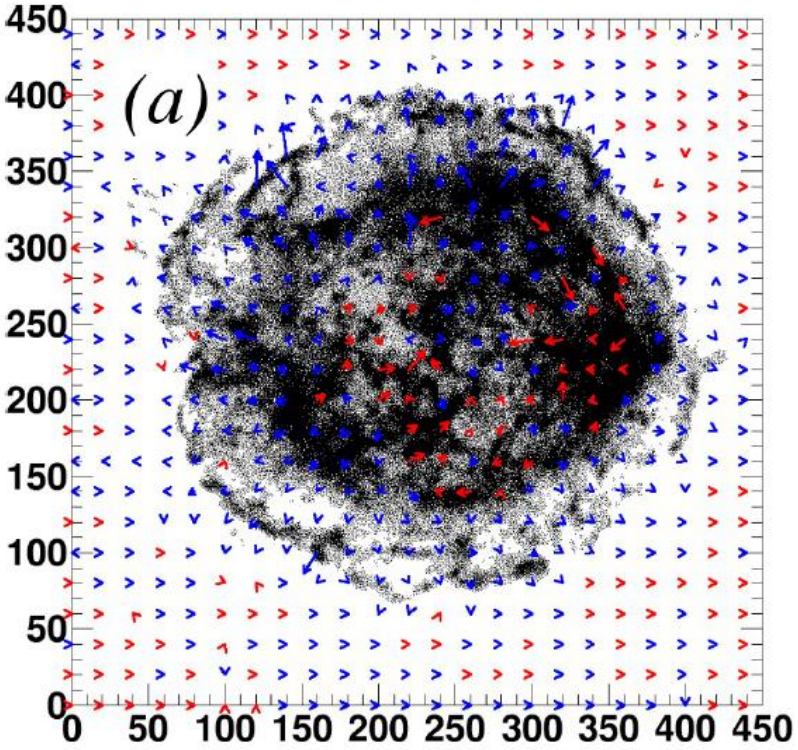
NuSTAR 衛星で取得した硬X線放射

Synchrotron 放射 (15-40 keV)

X線連続放射の逆行衝撃波領域と、硬X線放射領域が一致

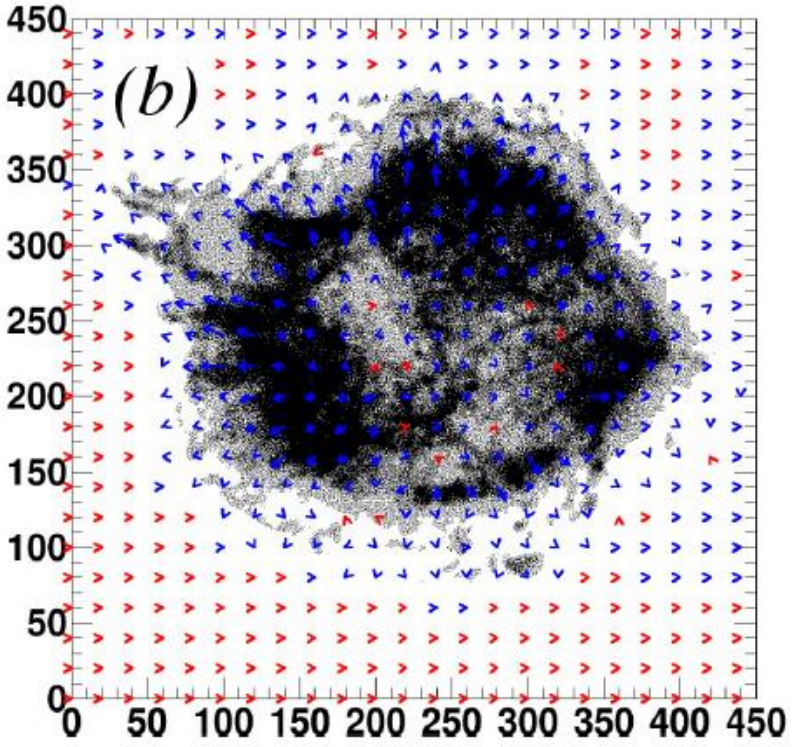
佐藤寿紀さん（立教大学）との共同研究

X線放射成分による速度場の違い



Chandra衛星 (2000年→ 2014年)

Synchrotron 放射 (4.1 – 6.3 keV)



Chandra衛星 (2000年→ 2014年)

Si輝線放射 (1.7-2.1 keV)

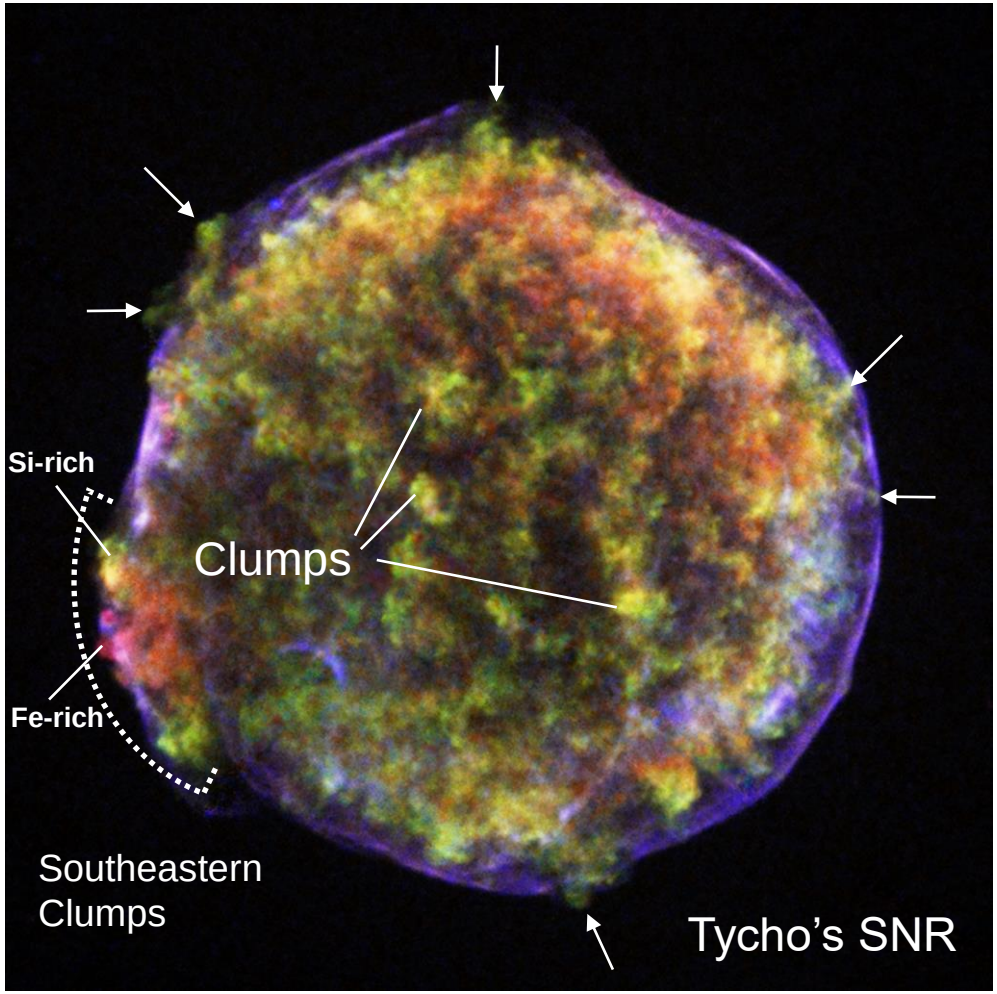
非熱的放射成分と熱的放射成分の動きが異なることが分かった。

Sato, et al. (2018), ApJ, 853, 46 “X-ray Measurements of the Particle Acceleration Properties at Inward Shocks in Cassiopea A”

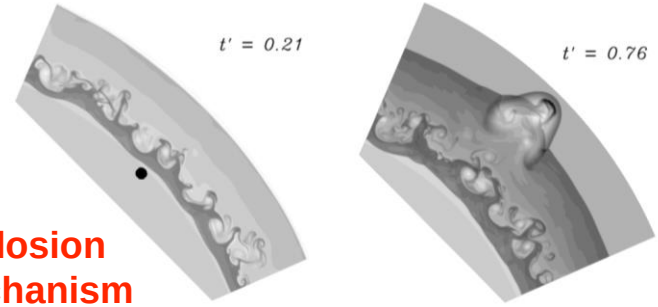
佐藤寿紀さん（立教大学）との共同研究

Ia型超新星残骸の塊構造の形成過程

塊構造(clumps) はどうやって形成されたか？

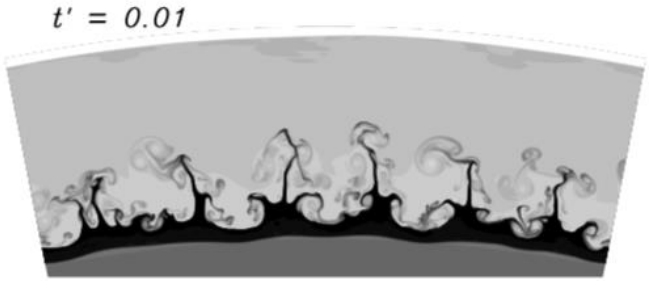


Initial clumps (^{56}Ni bubble)?



Explosion mechanism

Rayleigh-Taylor instability?



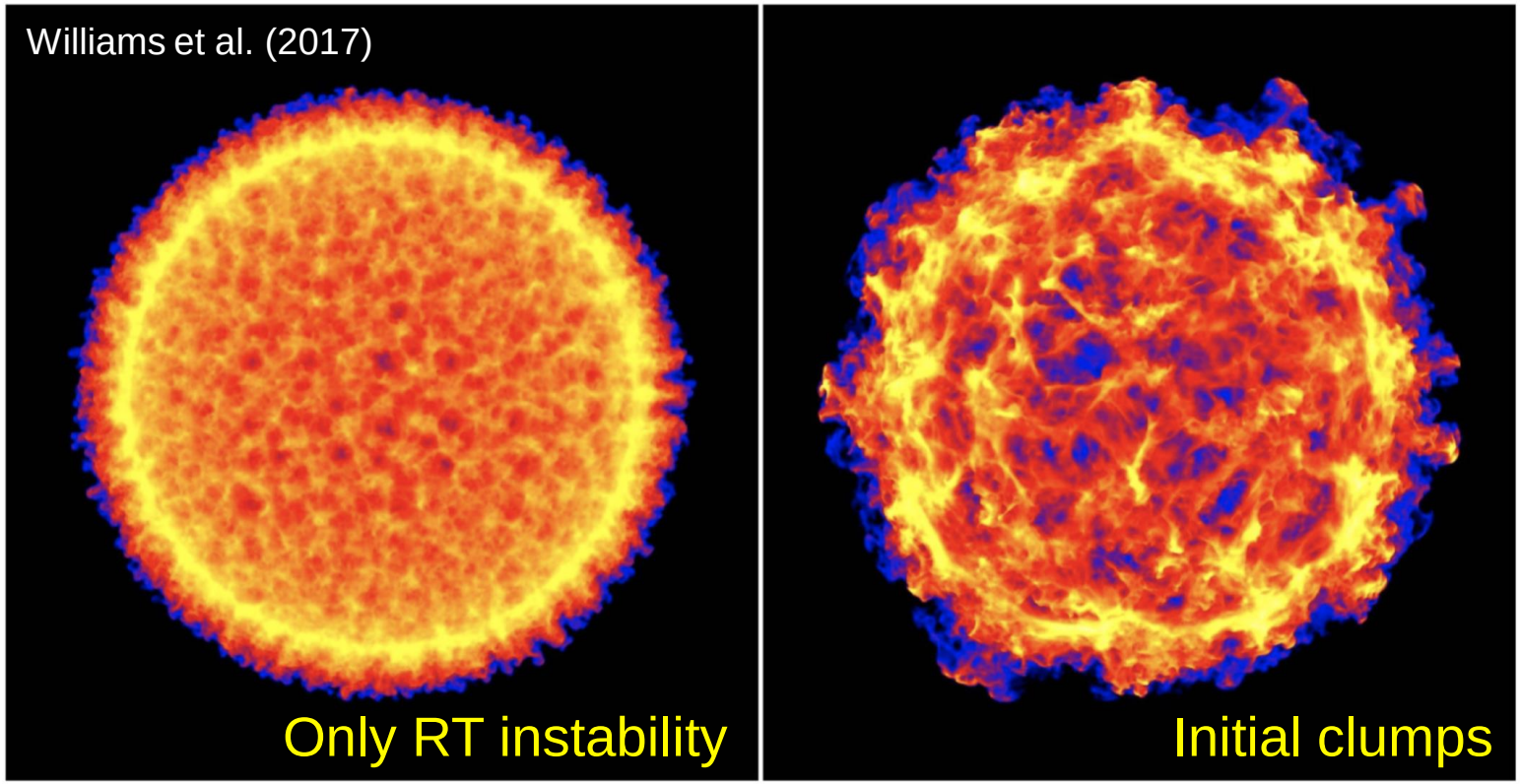
Hydrodynamics

e.g., Wang & Chevalier (2001), Warren & Blondin (2013)

This should be also related to the origin of the Fe-rich clumps!

佐藤寿紀さん（立教大学）との共同研究

Ia型超新星残骸の塊構造の形成過程



Initial condition

- Exponential density profile
- $E = 10^{51}$ erg
- $M = 1.4\ M_{\odot}$

Initial clump condition

"Smooth" noise
By a Perlin algorithm
(Perlin 1985)

- Max angular scale ~20°
- Max-to-min density contrast of 6

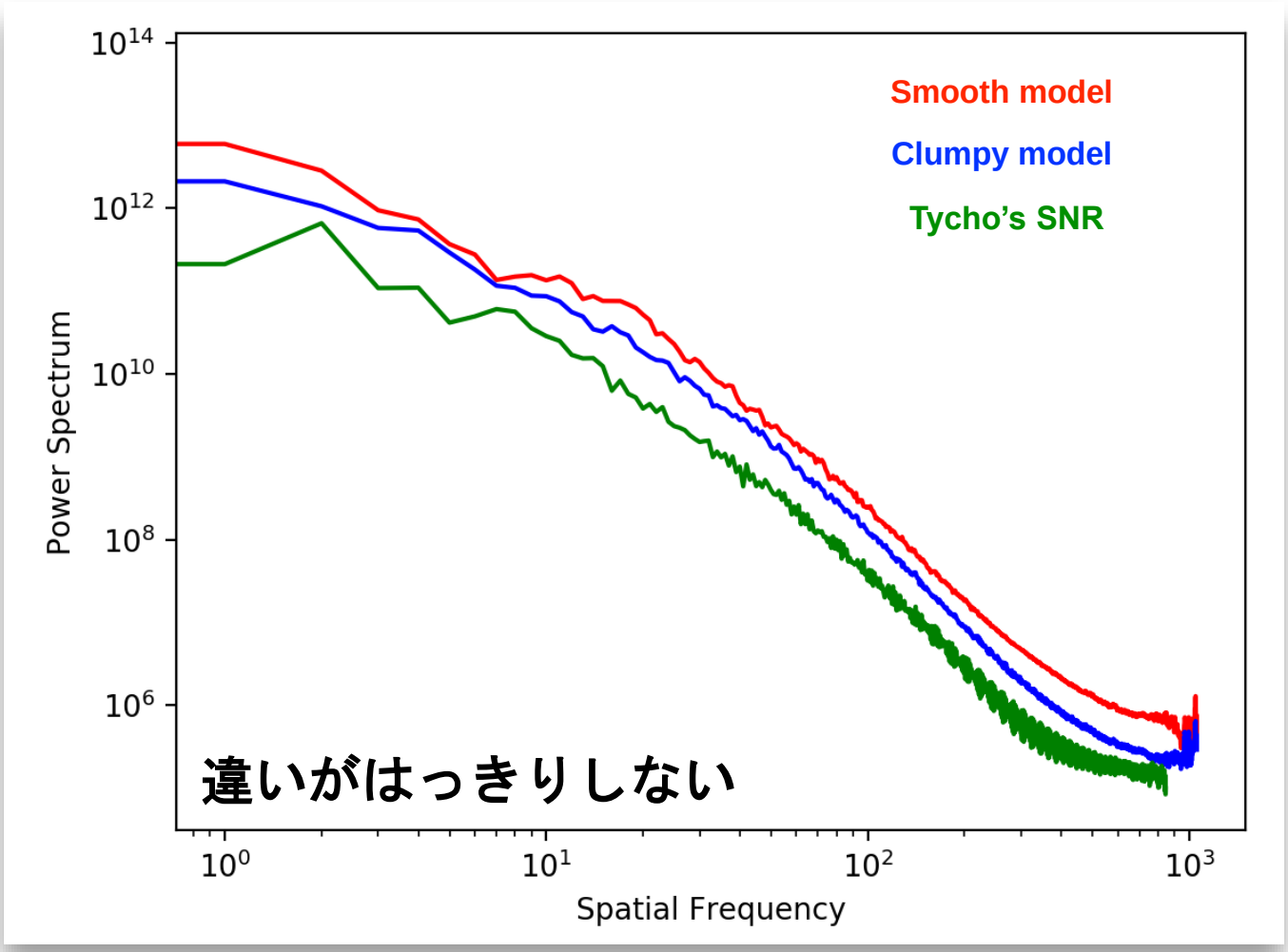
The clumpy structures seem to be much different from each other

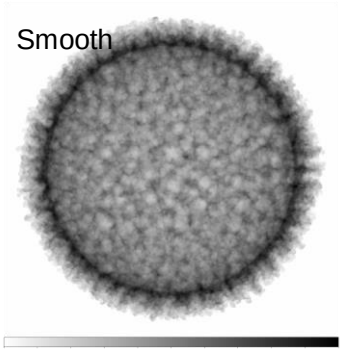
シミュレーションイメージとTycho SNRとを比較したい

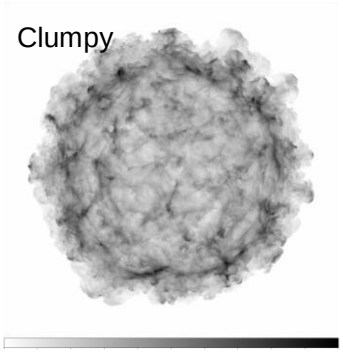
佐藤寿紀さん(立教大学)との共同研究

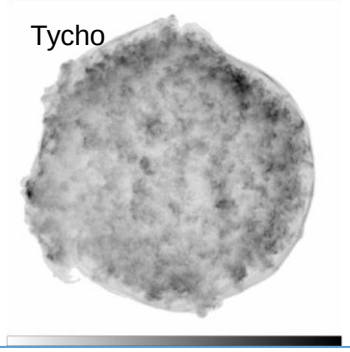
Ia型超新星残骸の塊構造の形成過程

2D Fourier transformation





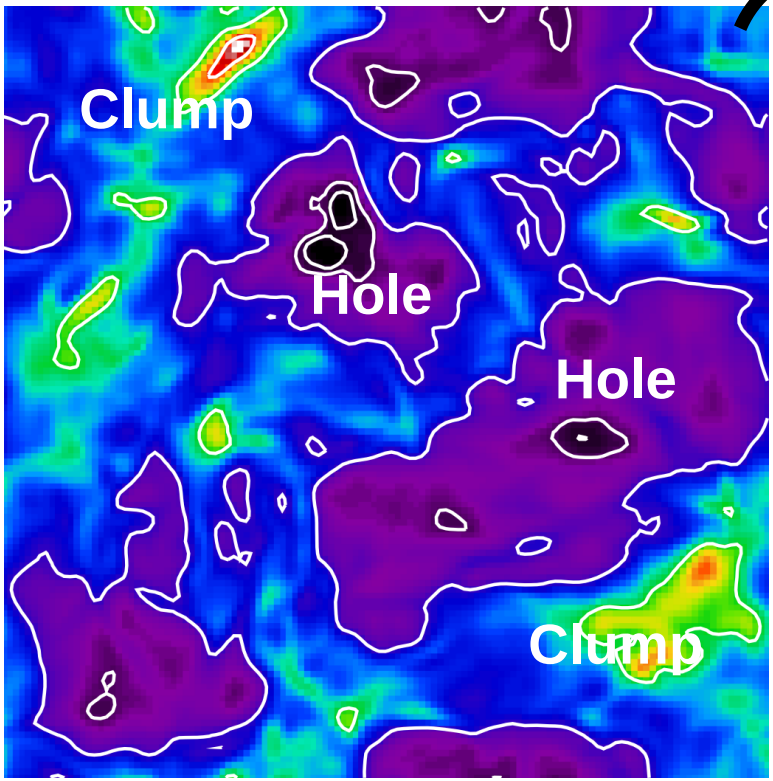


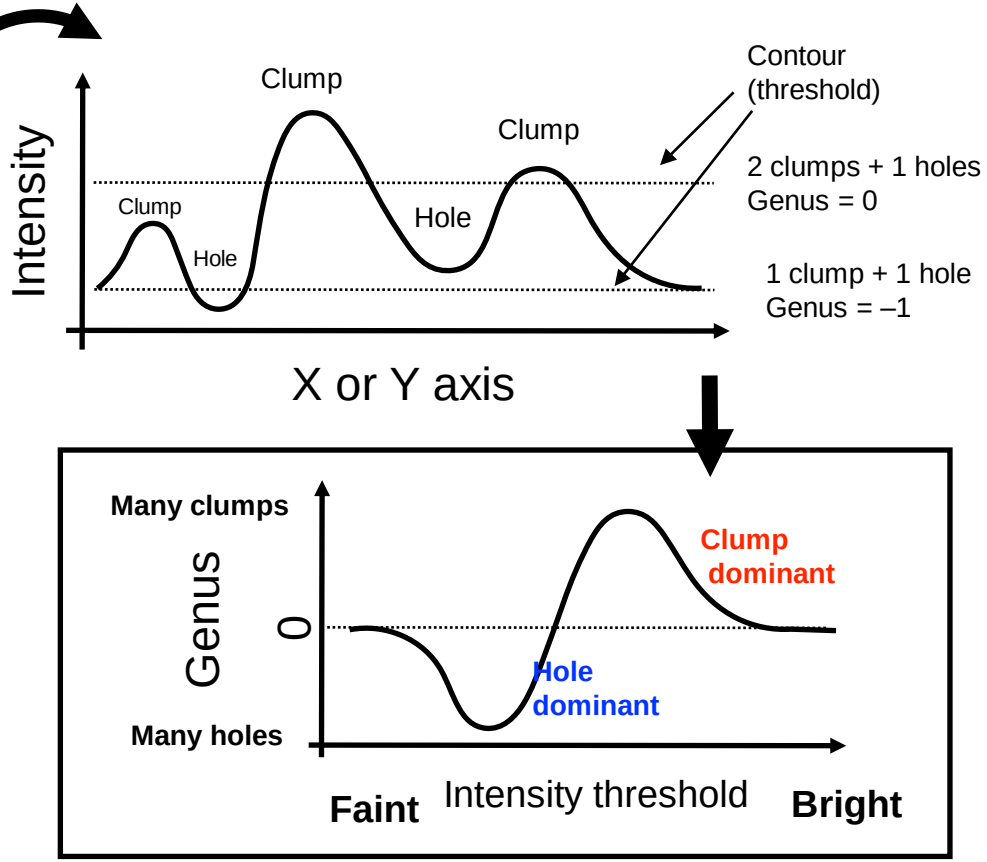


佐藤寿紀さん（立教大学）との共同研究

Genus Statistics

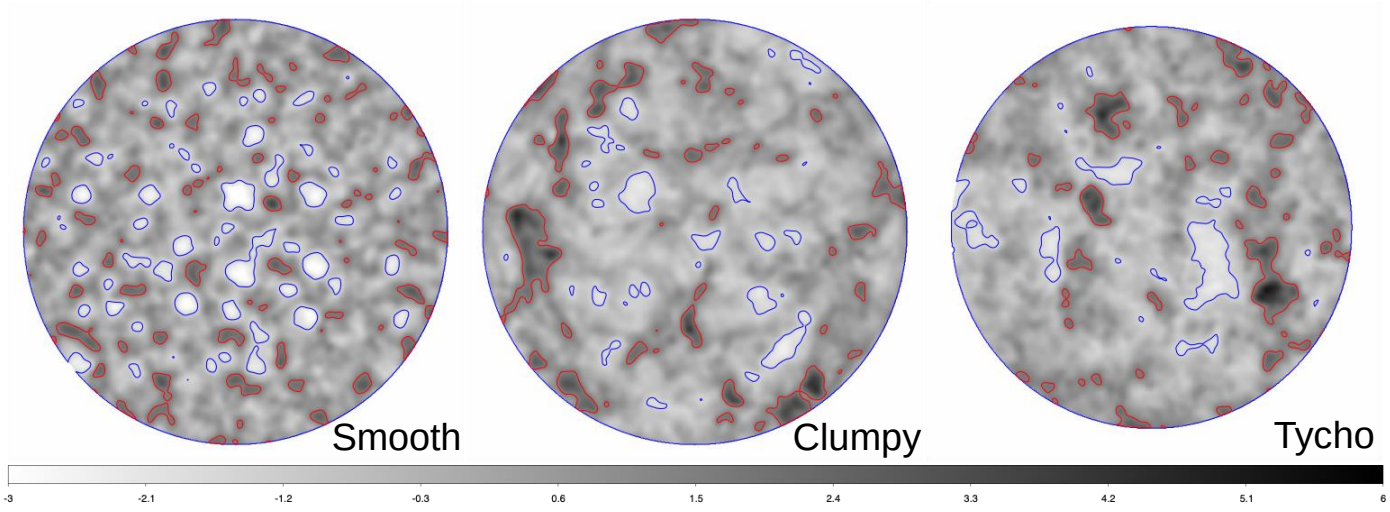
- 位相幾何学の概念。
- ある輝度等高線において：
- 　　オイラー標数（Genus 統計量）= (連結成分の個数) - (holeの個数)
- 宇宙の大規模構造の解析で使われている方法。





佐藤寿紀さん（立教大学）との共同研究

Comparison with hydro models



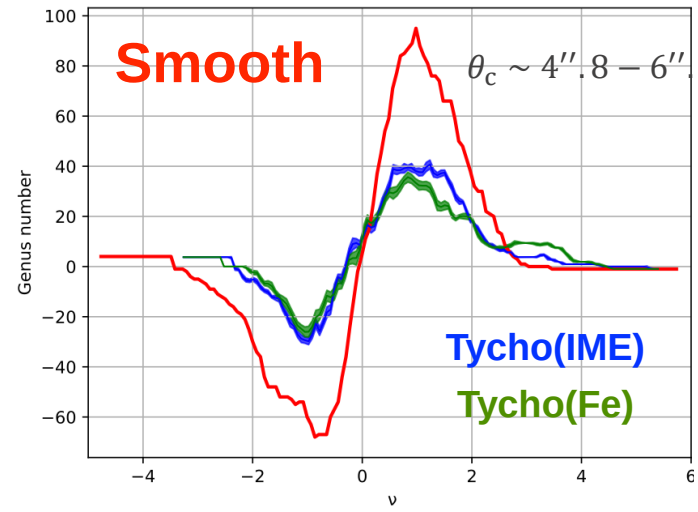
Smooth model

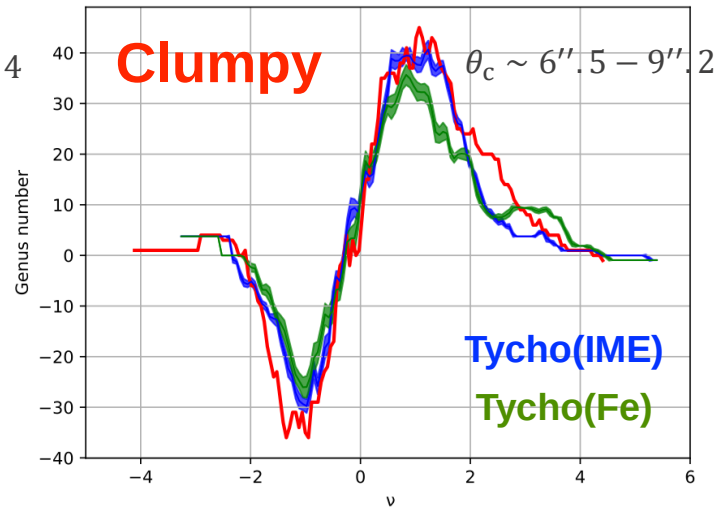
Clumps arise from the action of the dynamical instabilities acting through the duration of the remnant's evolution

- the clumps are homogeneously distributed across the central region
- the clumps are of similar shape and size

Genus curve → similar to

Gaussian distribution





Clumpy model

- the clumps are more filamentary
- the clumps are distributed more randomly

Genus curve → similar to

non-Gaussian distribution

→ similar to Tycho's SNR

Initial clumped ejecta distribution model が支持される。

佐藤寿紀さん（立教大学）との共同研究

企業で取り組んだこと:
データ科学の応用例

- ３例紹介します。
- 自動車関連企業の研究所からの依頼:
  - EUNN の論文実装。
  - FAB for HMEの論文実装。
- 某製造業を営む企業からの依頼:
  - 製品の設計過程を自動化、最適化したい。
  - 製造工程で、製品サイズが変化する様を、グラフで表現し、最短経路問題を用いて最適化した。

EUNNの論文実装

- 自動車関連企業の研究所からの依頼
- Efficient Unitary Neural Network(EUNN)の論文実装。
  - L. Jing et al. (2017), Proc. of ICML 2017, “Tunable Efficient Unitary Neural Networks (EUNN) and their application to RNNs”
- 論文を読んで解説して欲しい。
- 論文著者が公開しているTensorFlow 1 で実装されているプログラムを、TensorFlow 2に書き換えて欲しい。

EUNNの論文実装

- Recurrent Neural Network(RNN) :
  - 時系列データの予測でよく使われるDeep Learningの手法の一種。
- Efficient Unitary Neural Network(EUNN)はRNNのセルの一種。
- EUNNは、Unitary行列を用いることでVanishing/Exploding Gradientsの問題を防止できる。

EUNNの論文実装

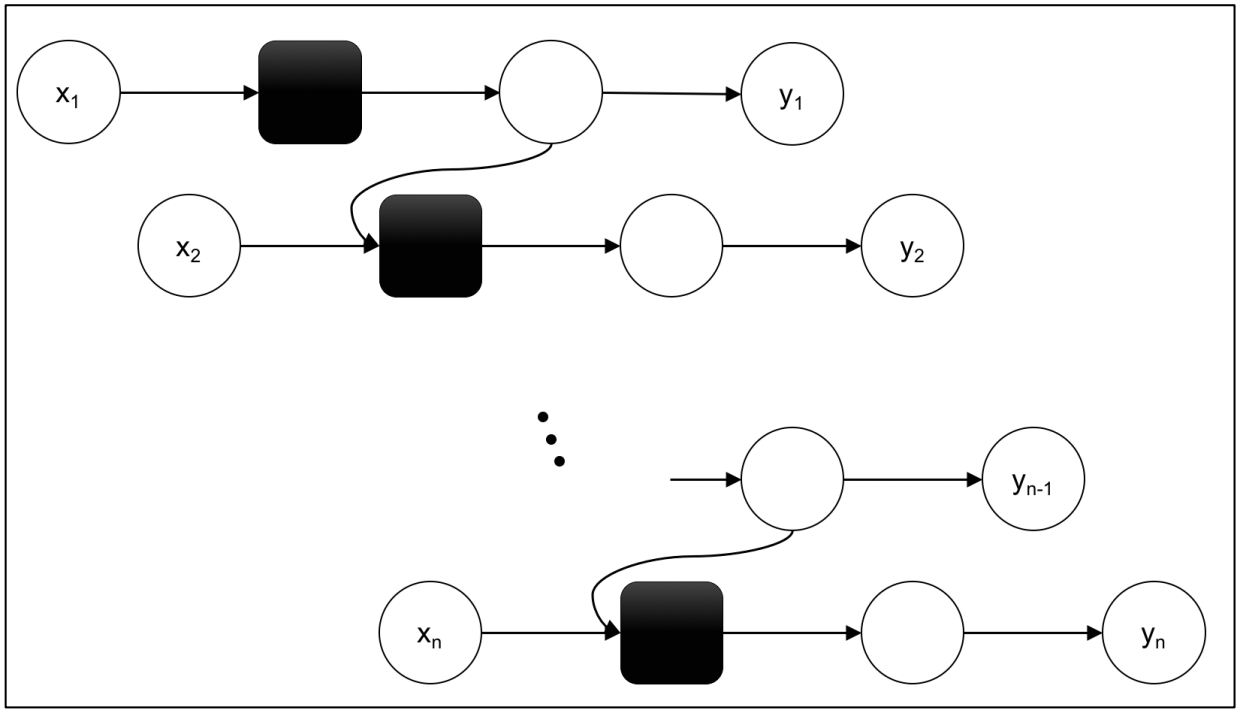
- Recurrent Neural Network(RNN)：　時系列データに対応したニューラルネットワーク。時系列のデータポイントは、各層の入力として利用される。また、各層の出力は、次の層の入力としてだけでなく、ユーザーが使用可能な出力としても利用される。

入力(x_1, x_2, …, x_n)
出力(y_1, y_2, …, y_n)

ある層の出力は、次の層の入力として利用される。また、出力としても利用される。

ある層の入力には、前の層の出力と、時系列のデータポイントを与える。

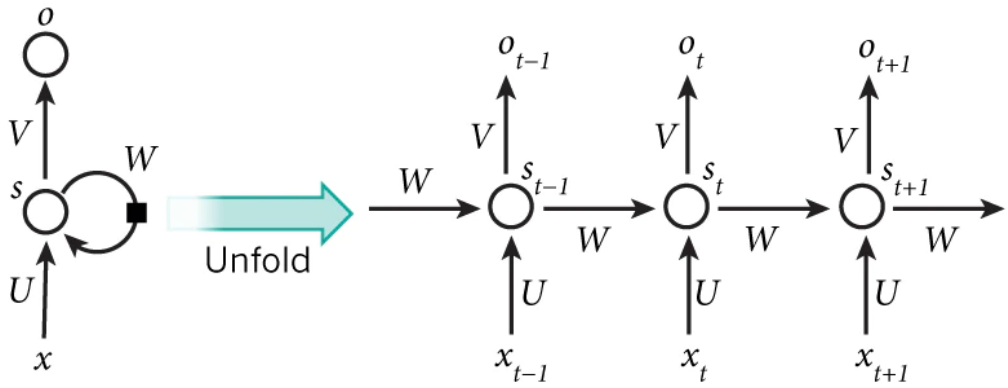
セル(右図の黒色部分):
RNNの隠れ層において、再帰的に出現する同一のネットワーク構造のこと



https://atmarkit.itmedia.co.jp/ait/articles/1804/25/news143.html

EUNNの論文実装

- Recurrent Neural Networks



① 入力：$x^{(t)}$ (t = 1, 2, …)

② Hidden layer: $h^{(t)}$ を、(1) 式で更新する。
※ σは、non-linear activation function

③ 出力：$y^{(t)}$ を(2) 式で出力する。

$$\mathbf{h}^{(t)} = \sigma(\mathbf{U}\mathbf{x}^{(t)} + \boldsymbol{W}\mathbf{h}^{(t-1)}), \qquad (1)$$

$$\mathbf{y}^{(t)} = \mathbf{W}\mathbf{h}^{(t)} + \mathbf{b}, \qquad (2)$$

EUNNの論文実装

- Vanishing and Exploding Gradient Problem
  - Cost function をgradient descent で更新するが、その時、以下の計算をすることになる。

$$\frac{\partial C}{\partial \mathbf{h}^{(t)}} = \frac{\partial C}{\partial \mathbf{h}^{(T)}} \frac{\partial \mathbf{h}^{(T)}}{\partial \mathbf{h}^{(t)}} \tag{3}$$

$$= \frac{\partial C}{\partial \mathbf{h}^{(T)}} \prod_{k=t}^{T-1} \frac{\partial \mathbf{h}^{(k+1)}}{\partial \mathbf{h}^{(k)}} \tag{4}$$

$$= \frac{\partial C}{\partial \mathbf{h}^{(T)}} \prod_{k=t}^{T-1} \mathbf{D}^{(k)} \mathbf{W}, \tag{5}$$

  - Wの乗算の部分は、Wの固有値が λ >> 1 だと発散し、λ << 1 だとゼロになってしまう。
  - Cost function を更新するときに必要なGradientが発散したり、ゼロになったりするため、RNNが上手く機能しなくなる。
- Unitary RNNs
  - W 行列を、Unitary 行列に制限すれば、固有値の大きさが 1 なので、上記の問題が発生しない。
  - Unitary RNNs がいくつか提案されている。
  - 本論文は、効率よくUnitary RNN の計算を行うアルゴリズム(Efficient Unitary Neural Networks)を提案している。

EUNNの論文実装

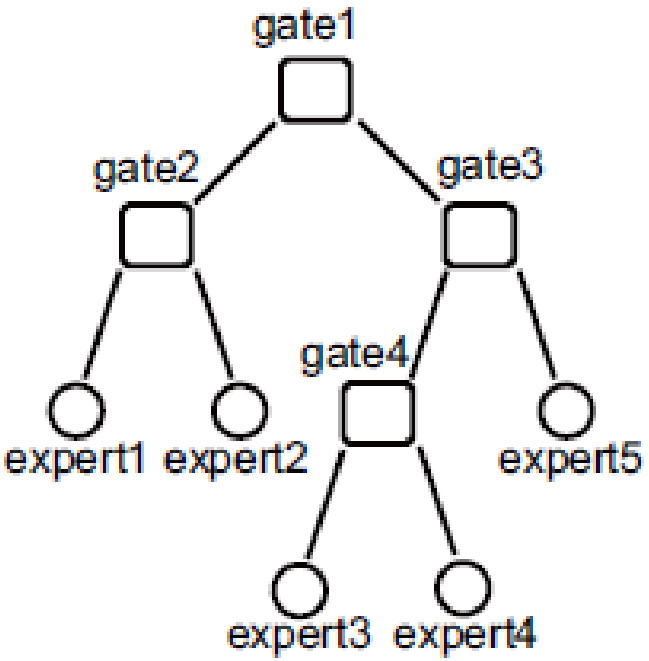
- 詳細は割愛します。
- 困ったところ:
  - 論文に記載されているアルゴリズムと、著者が公開しているソースコードでは、計算式が異なっていた。
  - ユニタリ行列の位相部分の割り当て方が異なっているためであると判明した。
- 企業の利点と思う点:
  - プログラミングスキルの高い社員が沢山居るところ。
  - 論文理解と解説は森井が担当した。
  - TensorFlow1 をTensorFlow2に書き換えるタスクについては、TensorFlowに慣れた社員が担当した。
  - うまく分業することで、課題をクリアすることができた。

FAB for HMEの論文実装

- 自動車関連企業の研究所からの依頼
- Factorized Asymptotic Bayesian (FAB) Inference for Hierarchical Mixtures of Experts (HMEs):
  - Fujimaki & Morinaga (2012), Proc. of AISTATS, “Factorized Asymptotic Bayesian Inference for Mixture Modeling”
  - Eto et al. (2014), Proc. of AISTATS, “Fully-Automatic Bayesian Piecewise Sparse Linear Models”
- 論文のアルゴリズムを実装して欲しい。

FAB for HMEの論文実装

- 線形関数を繋ぎ合わせてデータ全体をフィットする。
- Tree のモデル：領域をgateで分割し、各部分をExpertと呼ばれるモデル（今回は、線形関数）でフィットする。
- データにモデルをフィットする際に、FIC と呼ばれる指標を最大化する計算を行う。



FAB for HMEの論文実装

- 詳細は、ざっくり割愛します。
- 困ったところ:
- フィットの収束をモニターするためには、FICという数値を計算する必要があるが、その計算方法が明示されておらず、自分で導出する必要があった。
- Algorithm に記載されている式にタイポがあり、修正する必要があった。（そのまま実装したら、無限ループに陥った。）

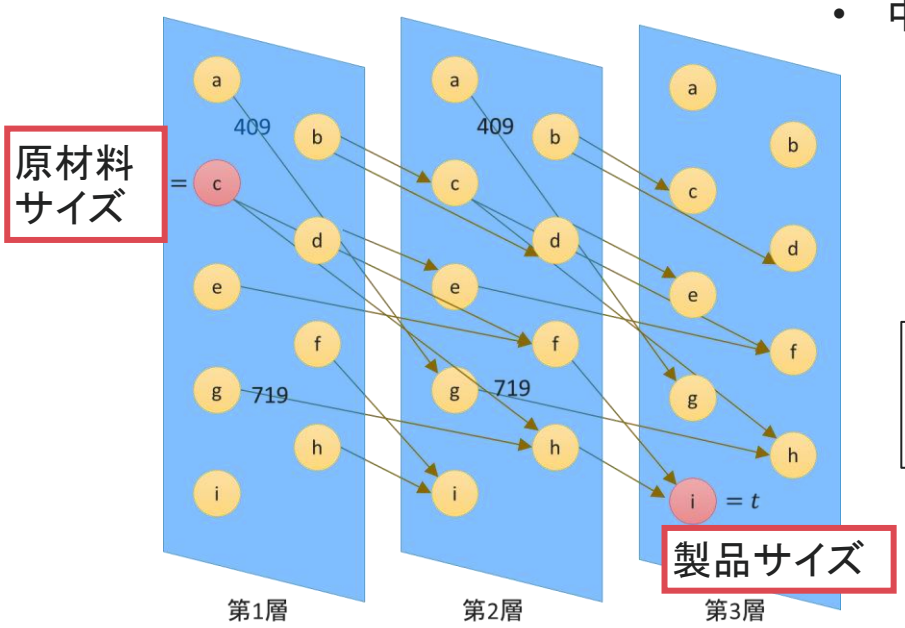
最短経路問題を用いた工程設計の最適化

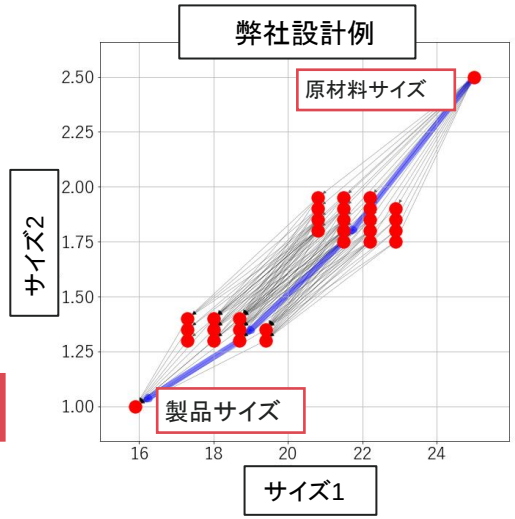
製造業を営む企業からの依頼：製品の製造工程を自動化・最適化して欲しい。

具体的な製品名などを用いて説明することが許されないため、イメージが湧きにくい説明になりますが、ご了承ください。

- 製造工程： 原料に対し複数回の加工を行って、製品サイズに加工する。
- 課題： 中間段階のサイズに任意性がある。最適な中間段階のサイズを見つけ出す。

- 解法： 中間段階のサイズを離散化し、加工工程を矢印で結んでグラフを作成し、最短経路問題として解く。
- 中間段階のサイズをグラフ上のノードとする。



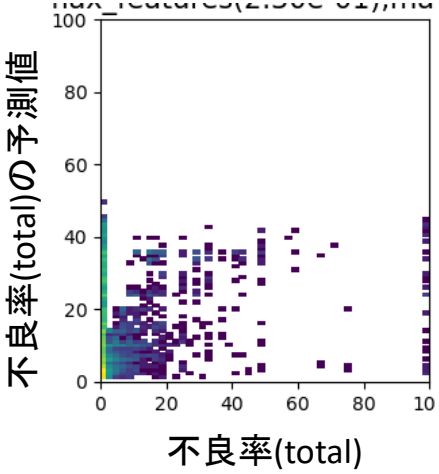


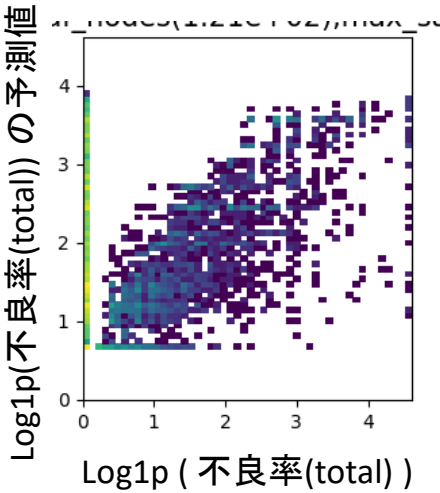
- 過去の製造実績データから作成した確率値をグラフのエッジに割り当て、確率を最大にする経路を求める。
- 左図青線が、最適な製造工程を示す。

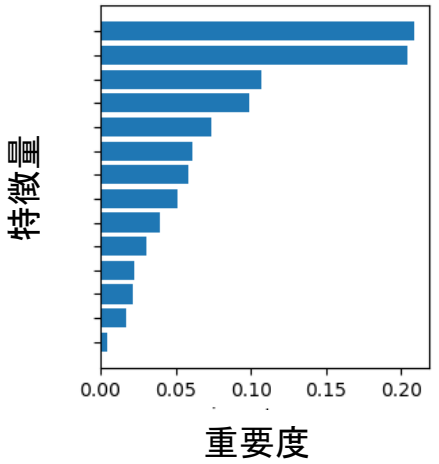
不具合情報の反映

- 過去の製造実績データから、加工工程の不良率を得ることができる。
- 加工工程における様々な条件（10数個程度)から不良率を予測するために、機械学習を用いた。
- 機械学習ライブラリを用いる際、ハイパーパラメータチューニングのために、ベイズ最適化を用いた。

- 手法:
- Random Forest 回帰
- ベイズ最適化を用いてハイパーパラメータチューニング
- 5-fold Cross Validation
- 不良率に対して、Log1p変換 log(x + 1) をして学習させる。
- 結果:
- 右上図で、Log1p(不良率) と、その予測値が一致する線上にデータが並べば、予測性能が良いことを意味する。
- 左下図は、予測に寄与する特徴量の重要度を示す。
- グラフ上のエッジに付与する重みとして、ここで得られた加工成功率を用いる。







最短経路問題を用いた工程設計の最適化

- 最短経路問題を解く部分は、python のライブラリ関数を用いて簡単に解ける。
- ダイクストラ法を用いた。計算時間も短い。
- しかし、案件自体は困難であった。その理由は、、、
- ノード間にエッジを引くときに、様々な制約条件がある。そのため、グラフを作成すること自体が難しい（面倒くさい）。
- 様々な規格の製品があり、製造方法が多数ある。それらに対応するために場合分けが多数必要になる。
- そのため、力技でコーディングする能力が必要であった（データ科学とは関係ない）。
- 依頼主からの聞き取り作業が多数発生。何度も何度もミーティングを実施（忍耐が必要）。
- クラスの設計、リファクタリング、コードの可読性、バージョン管理、テストといった地道な作業がほとんど。
- このようなスキルは、天文学の研究をしたときに得たもの。特にMAXIプロジェクトに参加していたときに得たもの。

企業におけるデータ科学

- 弊社におけるデータ科学の応用例を３例紹介しました。
- この中で、EUNN とFAB/HMEはデータ科学としては高度だが、収益性はあまりない。
- 最短経路問題を用いた製造工程の最適化のタスクは、データ科学としては簡単だが、一番収益性が高い。
- データ科学よりは、力技的なプログラミングスキルが必要であった。
- 顧客企業様からは大変感謝された。

- データ科学としては、簡単な方法を知っていれば十分。それらを上手く組み合わせることが必要。
- 実際の問題に対応するためには、データ科学はほとんど関係ないのかもしれない?

ポスドクから民間企業へ移って分かったこと

- 企業に就職してもさほど生活に大きな違いは感じていません。
- スーツは着なくてよい。職場が虎ノ門ヒルズのようなキラキラした場所でも、気にせず私服です。
- 客先（クライアント）に訪問したときも、スーツは着なくてもよかった。
  - 初めて天文学会、物理学会に参加した時、研究室の先輩から「スーツ着ないとだめだよ」と言われましたが。
- 社員は、学生時代の専門がバラバラなので、お互いに尊重し合う雰囲気がある。
- Data Scientist, エンジニア, ビジネスマンは随分異なる。
- ポスドクの時とは違って、収入は安定している。任期切れの心配がない。
- IT企業はコロナ禍でも増収益なので余裕がある。リモートワークには、簡単に対応できた。現在も継続中。
- 企業であっても、こき使われることはない。自由な時間はある。数学の勉強も進みました。
- 天文学の研究で習得したスキルは、他の分野でも使えると思います。

END

概要

- 本講演者は、X線天文学で博士を取得した後、全天X線監視装置「MAXI」プロジェクトに参加して時間軸天文学に貢献する成果を挙げた。その後研究分野を変え、統計数理研究所にて観測天文学にデータ科学の手法を応用する研究を行った。2019年よりデータ科学をビジネスに応用するIT企業に勤めている。本講演では、天文学のデータ解析にデータ科学を応用した例をいくつか紹介する(超新星残骸の成分分離、画像解析に関する研究(Katsuda et al. 2018, Sato et al. 2018, Sato et al. 2019)。また、天文学データではないが、IT企業におけるデータ科学に関する仕事の例も紹介する(EUNN の実装、FAB for HMEの実装、最短経路問題を用いた製造工程の最適化)。講演者は、結局アカデミック業界からビジネス業界に転身したが、データ科学という共通言語を用いることによって、ビジネスにおける問題解決においても面白い仕事ができることをお伝えしたい。